The Relationship between Accuracy and Interval Length in the Respondent-Generated Interval Protocol

S. James Press
University of California at Riverside, Riverside California 92521-0138

Judith M. Tanur
State University of New York at Stony Brook, Stony Brook, New York 11794-4356

KEY WORDS: RGI, Bayesian Estimation, Brackets, Cognition, Survey Design

Introduction

In the Respondent-Generated Interval (RGI) technique (see Press, 2003), respondents are asked to recall a numerical quantity by giving both a best estimate and an interval that bounds the values that the respondent thinks the recalled quantity might take. The RGI technique then estimates a population mean using a weighted average of the values given by the respondents. The weights are functions of the intervals supplied by the respondents; longer intervals, assumed to represent less accurate recall, generate smaller weights than do shorter intervals, assumed to represent more accurate recall.

Press (2003) has shown, through some toy examples, that as long as accurate respondents give short intervals and inaccurate respondents give longer ones, the RGI estimator is less biased (in the sense of containing less respondent error) than the sample mean. This finding held even in the case where 99 hypothetical respondents made inaccurate best guesses and gave long intervals while only one hypothetical respondent made an accurate best guess and gave a short interval. The current paper explores whether accurate real respondents using an appropriate form of the RGI bounds questions do indeed tend to give shorter intervals than do inaccurate ones. It ends by proposing that a form of the RGI bounds question be designed to increase the correlation between accuracy and interval length.

To carry out this exploration we first examine data generated in the spring of 1997 when approximately 1000 students at the University of California at Riverside (UCR) and about 750 at the State University of New York at Stony Brook (SUSB) responded to a paper-and-pencil, self-administered questionnaire asking them to recall factual information about matters relating to their life on campus using the RGI protocol. (See Press and Tanur, 2000 for a fuller description of these experiments.) At both campuses students were asked for the number of credits they had earned (CREDITS), the number of grades of C or less they had received (C's), their grade point average (GPA), their Scholastic Aptitude Scores on the math (SATM) and verbal (SATV) tests, and the number of traffic tickets they had received on campus during that academic year (TICKETS). At UCR students were also asked to recall the amount of the registration fee (REGFEE) and the recreation fee (RECFEE) they paid at the beginning of the quarter. The corresponding questions at Stony Brook asked for the amounts of the health fee (HEALTH) and the student activities fee (SAFEE). Stony Brook students were also asked to recall the amount spent on the food plan (FOOD) and the number of library fines (FINES) they had been assessed.

For those respondents who consented to have their data verified and who gave their student identification numbers for that purpose, we were able to get "true" values for the usage quantities the students were recalling from appropriate campus offices. (No permission for verification or identification numbers were needed to ascertain the true values for the fee data, as the fees are standard for all full-time students. We limited our analysis to full-time students.) While we are cognizant of the possible errors in the administrative data we are using for verification, we shall use those data as a "gold standard" in what follows.

The sample sizes vary across questions for several reasons. First, for most of the items the question was asked in the interval form described here for only half the respondents, as we had originally thought to test another form of the bounds question and hence used a split ballot design. (We gave both question forms of the fees question, so the entire sample answered in the form we were interested in; hence the sample sizes are larger for those questions.) Second, not all respondents gave permission to verify their data. (Because no verification was needed for the fee data and hence finding "truth" was not contingent of a respondent's permission, the sample sizes for the fee data remained large.) Third, some respondents were not eligible to answer some questions – e.g. some had not taken SATs and some were not on the food plan. Finally, the records of the verifying sources varied in quality, and some sources were unable to locate records for some of our respondents who supplied ID numbers.

Implications of the Campus Experiments

We used the difference between the upper bound and the lower bound offered by respondents to measure length of interval. Accuracy was measured as the absolute value of the difference between the respondent's best guess and truth as reported by the verification office. (Note that this is actually a measurement of **inaccuracy** rather than of accuracy, so the negative correlation we hope to find between accuracy and interval length will be reflected in a positive correlation between this measure of absolute error and interval length.) Table 1 shows the product moment correlations between interval length and absolute error for both UCR and SUSB students. We see that the relatively large correlations between interval length and absolute error are for number of grades less than C and number of traffic tickets at both campuses, GPA at UCR, and number of library fines at SUSB. With the exception of GPA at UCR, these are all socially undesirable items. Significance levels are supplied as a matter of interest, even though the data are not normally distributed.

Table 1—Correlations between length of interval and accuracy
USING ALL DATA

	UCR			**SUSB**		
VARIABLE	**n**	**r**	**sig**	**n**	**r**	**sig**
Credits	129	-.02	.827	132	.07	.435
C's	118	.57	<.001	131	.30	<.001
GPA	131	.49	<.001	141	.08	.334
SATM	102	.23	<.020	81	.15	.187
SATV	97	.13	.193	82	.20	.074
Tickets	137	.61	<.001	139	.50	<.001
RegFee	656	-.07	.064			
RecFee	700	.22	<.001			
SAFee				426	.07	.145
Health				444	.12	.010
Food				68	.22	.066
Fines				126	.92	<.001

.

An examination of the scattergrams showed that there were often outliers in the length distributions (perhaps generated by respondents who were not taking the interval generating task very seriously), and these were often influential points in the scattergrams. Hence we re-ran the correlations deleting outliers on the interval length variables. For these purposes we defined (along with Moore, 2000, p. 74) an outlier as any observation that fell beyond 1.5 times the interquartile range above the third quartile of the distribution.[1] The correlations without outliers are shown in Table 2. Here the pattern is similar to that found in Table 1, with the correlations between interval length and absolute error remaining relatively large for the two socially undesirable items common to both campuses (number of grades less than C and number of traffic tickets). The correlation for library fines at SUSB, highly influenced by a single outlier, was substantially reduced, but still positive and statistically significant. The correlation for GPA at UCR dropped somewhat, while the correlation for verbal SAT score became substantial at SUSB as did that for math SAT at UCR. While some of these correlations are indeed relatively large, none of them accounts for more than 35% of the variance in accuracy, and most account for far less.

Because we were suspicious about the distributional properties of the variables

[1] There were two exception to this rule. For tickets at SUSB 5 clear outliers did not meet the criterion outlined in the text but these were matched by no accuracy measurement, so were not included in either correlation. Fines at SUSB had an IQR and a Q3 of 0, but values over 20 were clear outliers and were eliminated.

Table 2—Correlations between length of interval and accuracy
OUTLIERS DELETED

	UCR			SUSB		
VARIABLE	**n**	**r**	**sig**	**n**	**r**	**sig**
Credits	116	.00	.958	124	-.04	.650
C's	108	.59	<.001	125	.40	<.001
GPA	121	.38	<.001	137	.15	.087
SATM	102	.23	<.020	76	-.02	.836
SATV	90	.11	.317	79	.34	.002
Tickets	130	.49	<.001	139	.50	<.001
RegFee	618	-.25	<.001			
RecFee	651	.07	.095			
SAFee				407	-.03	.574
Health				407	-.05	326
Food				63	-.14	.268
Fines				124	.29	.001

involved, we decided to calculate rank order correlations between interval length and absolute error as well. These results are shown in Table 3 (for all data) and Table 4 (with outliers removed).

Table 3— Rank Order Correlations between length of interval and accuracy
USING ALL DATA

	UCR			SUSB		
VARIABLE	**n**	**r**	**sig**	**n**	**r**	**sig**
Credits	129	.28	.002	132	.26	.002
C's	118	.73	<.001	131	.37	<.001
GPA	131	.37	<.001	141	.14	.088
SATM	102	.34	.001	81	.17	.122
SATV	97	.41	<.001	82	.23	.037
Tickets	137	.57	<.001	139	.39	<.001
RegFee	656	-.14	<.001			
RecFee	700	.16	<.001			
SAFee				426	-.15	.002
Health				444	-.09	.063
Food				68	.05	.672
Fines				126	.46	<.001

questions that was used in the campus experiments resulted in correlations that were not We see the same pattern appearing, with the correlations at SUSB smaller than those at UCR, but in both cases largest correlations appearing for the socially undesirable items of number of grades less than C, number of traffic tickets, and number of library fines. It seems that using the form of asking the bounds high – not nearly as high as that produced in the toy examples of Press (2003), which ranged from .624 up to 1. We need to use a form of asking the bounds questions that results in a stronger relationship between accuracy and interval length to ensure the success of the RGI protocol in reducing bias.

Table 4— Rank Order Correlations between length of interval and accuracy
OUTLIERS DELETED

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

References

Moore, David S. 2000, *The Basic Practice of Statistics, 2nd Edition.* New York: W.H. Freeman and Company.

Press, S. James, 2003, "Respondent-Generated Intervals for Recall in Sample Surveys," available at www.statistics.ucr.edu/press.htm (This is the basic reference that gives the derivation of the RGI procedure.)

Press, S.. James and Judith Tanur, 2000, "Experimenting with Respondent-Generated Intervals in Sample Surveys", with discussion, in *Survey Research at the Intersection of Statistics and Cognitive Psychology,* Working Paper Series #28, Monroe G. Sirken, Editor, National Center for Health Statistics, Jan. 2000, U.S. Department of Health and Human Services, Center for Disease Control and Prevention, pp. 1-18. (Note that the model used to generate estimates for the empirical work in the Press/Tanur 2000 paper has been superceded by the model described in Press, 2003, cited above.)