RESPONSE ERROR EFFECTS OF SURVEY QUESTIONNAIRE DESIGN

Gad Nathan, Hebrew University
Monroe G. Sirken, National Center for Health Statistics

1. Introduction

The questionnaire design (QD) is widely recognized as the weak link in the survey measurement process. That QD effects occur and can be substantial has been documented in many quality check and pilot studies, but as noted by Sirken (1986) and many others, the effects are rarely measurable in the main surveys. Except for the precautions that are exercised in developing and testing survey instruments, the error effects of QDs on survey statistics are largely uncontrolled and have been in that unfortunate state for a long time. One long range strategy for improving the art of designing questionnaires is to gain a better understanding of the role of cognition in the production of survey response errors and to incorporate this understanding in developing improved methods of estimating QD error effects.

This paper presents a statistical theory for measuring response errors due to QD effects in sample surveys that utilize multiple QD versions. The theory was developed as part of a research program in cognition and survey measurement that is being conducted at the National Center for Health Statistics with the support of a research grant from the National Science Foundation. The objectives and rationale of this multi-discipline research program were described by Lessler and Sirken (1986). In brief, the program proposes to stimulate interdisciplinary research on the cognitive aspects of QD, and on statistical methods for measuring the QD error effects.

The model presented in this paper is based on the classical test theory of Lord and Novick (1968), and the response error models of Hansen, Hurwitz, and Bershad (1961). It decomposes the total variance into the QD variance and the residual variance. The latter is the sampling variance; the former is the variance of the expected responses over a universe of exchangeable QDs with the following properties:

Property 1 - the universe contains two or more interchangeable versions of the questionnaire, differing in design but with the same data objectives.

Property 2 - the expected value of responses over all interchangeable QDs is unbiased.

The model assumes a survey is designed to estimate P, the proportion of positive responses to a dichotomous variable. The survey is based on a sample of "k" QDs, selected at random from the universe of exchangeable QDs, which is then randomized over a random sample of "n" reporting units. Estimates of P, and of the sampling and the QD variances are derivable from the survey data if $k>1$. (The model implies that the QD variance effects exist though they are unmeasurable from the survey data if $k=1$.) The model could be readily adopted in most surveys since the QD variance estimator does not require non-survey data sources of validating information. Furthermore, the model provides a strategy for controlling the QD error effects by varying "k" and for controlling the joint effects of sampling and QD variance errors by optimizing "k" and "n". Neither of these possibilities, however, is expressly addressed in this report.

Traditionally, the QD error effects have been expressed in terms of the systematic response errors. The QD model, on the other hand, proposes to measure QD effects in terms of the variation in the expected responses among the universe of interchangeable and biased QD versions. Given the model's advantage in being able to estimate the QD error effects without reference to non-survey sources of validating information, what evidence is there to justify such a change in viewpoint?

The change in perspective is compatible with the findings of Marquis, Marquis, and Polich (1986). After reviewing the literature for information about the direction and size of the response biases in sensitive topic surveys, they observed that the distribution of the biases seemed to center on zero rather than on negative values. The finding leads them to the conclusion that the attention that heretofore was given to controlling under-reporting biases in sensitive topic surveys could more appropriately be redirected to finding ways of reducing response unreliability.

The practical utility of applying the QD model depends on the feasibility of constructing QD universes for which the expected value over responses to all interchangeable QD options is unbiased. Constructing QD universes with the unbiasedness property is more likely to apply in some situations than in others. It would appear to be most feasible and most needed in situations where it is known or suspected that particular QD features such as the ordering of questions or response categories, using unbounded recall periods, etc., are likely to produce systematic cognitive errors in the respondents' judgments due to conditioning, anchoring, telescoping, etc. In situations such as these, where the responses to a single QD version are likely to be biased, it may be feasible to construct a "symmetric" universe of QDs such that the expected value of responses over all QDs would be unbiased because it averaged out the systematic errors associated with each QD version.

That the average response to several QD versions may be subject to smaller biases than the response biases of a single QD version appears to have been largely overlooked, but it is not a brand new idea. For example, the idea occurred to Monsees and Massey (1979), more or less accidentally, in the analysis of a split panel test of the "anchoring phenomena". The test investigated the effect on the reported income distribution of the order in which the income categories are asked in a telephone survey where it would be infeasible to use a flash card. The test was based on three QD versions which varied by the income category that was asked first. Each tested QD version yielded an income distribution that was biased toward the initial income category. However, the income distributions that were based on the average responses to two or three QD versions appear to be essentially unbiased.

Since the statistical properties of survey statistics based on the average responses to multiple QD versions have received little attention in the past, the feasibility of constructing unbiased QD universes is largely unknown and needs to be demonstrated by validity checks, like the one conducted by Monsees and Massey, or like those described by Jabine and Rothwell (1970) involving re-interviews or record checks.

A better understanding of the cognitive processes that lead to biased judgments would improve the prospects of constructing unbiased QD universes, and might ultimately provide a scientific underpinning to the art of designing questionnaires. Achieving these goals would be enhanced by improving communication between cognitive scientists and survey researchers and by applying the findings and methods of cognitive science in QD research. For example, the experimental findings of cognitive psychologists, Tversky and Kahneman (1974), on the heuristics that sometimes lead people to make biased judgments, as well as the experiments of Brown, Rips and Shevell (1985), on the cognitive processes that lead to biases in the temporal dating of past events, are highly relevant to problems in QD research.

In the absence of information about the unbiasedness of the QD universe, the model's estimator of QD variance could serve as a measure of response sensitivity to variations in QD versions. When large, the estimated QD variance would serve as a warning signal of response instability. When small, it would provide a degree of reassurance about the stability of response. The use of the QD variance in this way would be particularly suitable in attitude surveys since subjective phenomena are only accessible to respondents and cannot be validated by reference to external sources.

The remainder of this paper is devoted to the presentation of the QD model and illustrations of its application with existing data. The simpler of two versions of the QD model is presented in Section 2. The basic model, which assumes a single QD factor, is extended to the multivariate case in Section 3. Applications of the basic model are provided in Section 2 and of the extended model in Section 4. Some of the models' implications are briefly discussed in Section 5.

2. The Basic Model

The basic random effects model is developed in the following under fundamental simple assumptions, in order to form the basis for more complex models to be applied in practice. The model assumes that the variable of interest is a single dichotomous response variable, Y, and that the overall proportion of positive responses, P, is to be estimated. A simple random sample with replacement of size n is assumed. The response for any given sampled unit can be elicited by means of different design options, such as variations of questionnaires or ordering of questions. It is assumed that k options are selected randomly from a hypothetical large universe of possible design options so that their effects may be considered as exchangeable.

Let Y_{ij} be the observation for the j-th sample unit (j=1,...,n), if elicited by the i-th design option (i=1,...,k). We assume that

$$Y_{ij} = P_i + e_{ij} = P + D_i + e_{ij}, \quad (i=1,\dots,k;\ j=1,\dots n) \quad (1)$$

where P_i are i.i.d. random variables with $E(P_i) = P$ and $V(P_i) = V(D_i) = \sigma_D^2$, such that $Y_{ij}|P_i \sim B(1,P_i)$. Then, since unconditionally $Y_{ij} \sim B(1,P_i)$, we have that e_{ij} are i.i.d. with $E(e_{ij}) = 0$ and $V(e_{ij}) = P(1-P) - \sigma_D^2 = \sigma_E^2$ and $\mathrm{cov}(D_i, e_{ij}) = 0$ (although D_i and e_{ij} are not independent). Thus the total variance of Y_{ij} can be broken down as follows:

$$V(Y_{ij}) = P(1-P) = V(D_i) + V(e_{ij}) = \sigma_D^2 + \sigma_E^2, \quad (2)$$

where σ_D^2 is the design-option variance and σ_E^2 is the residual error (sampling) variance.

Since each subject can be observed via only one design option it is assumed that the k design options are allocated at random, so that each option is allocated to m = n/k sample units, assuming that n is a multiple of k. Let $\hat{P}_i$ be the proportion of positive replies (Y_{ij} =1) for the m units for which option i is used, so that

$$E(\hat{P}_i|P_i) = P_i \text{ and } V(\hat{P}_i|P_i) = P_i\ (1-P_i)/m. \quad (3)$$

Thus $\hat{P} = k^{-1} \sum_i \hat{P}_i$ is an (unconditionally) unbiased estimator of P, which is assumed to be the true value of the parameter of interest-- i.e., we assume no response bias.

The (unconditional) variance of $\hat{P}_i$ is obtained as

$$V(\hat{P}_i) = V[E(\hat{P}_i|P_i)] + E[V(\hat{P}_i|P_i)]$$
$$= V(P_i) + m^{-1}\ [P(1-P) - \sigma_D^2]$$
$$= m^{-1}\ P(1-P) + (1-m^{-1})\sigma_D^2$$

Thus:

$$V(\hat{P}) = \frac{P(1-P)}{n} \quad \frac{m-1}{n}\ \sigma_D^2$$
$$= \frac{P(1-P)}{n}\ [1+(m-1)\delta], \quad (4)$$

where $\delta = \sigma_D^2/[P(1-P)] = \sigma_D^2/[\sigma_D^2+\sigma_E^2]$, is the proportion of design option variance out of the total unit variance. The factor $1+(m-1)\delta$ therefore measures the increase in the variance of the estimator due to design-option variance (as compared to the situation where each unit is assigned to a different design-option, i.e., k=n and m=1) and may be termed the "design-option effect."

The components of variance can be estimated unbiasedly if (and only if) there are several design options, i.e., k>1. Then it is easy to show that

$$v(\hat{P}) = \frac{1}{k} S_P^2 = \frac{1}{k(k-1)} \sum_i (\hat{P}_i - \hat{P})^2 \qquad (5)$$

is an unbiased estimator of the variance of the estimator (4) and that

$$\hat{\sigma}_D^2 = \frac{1}{n-k} [(n-1)S_P^2 - k\hat{P}(1-\hat{P})] \qquad (6)$$

and

$$\hat{\sigma}_E^2 = \frac{n}{n-k} [\hat{P}(1-\hat{P}) - \frac{k-1}{k} S_P^2] \qquad (7)$$

are unbiased estimators of σ_D^2 and of σ_E^2, respectively. Note that $\hat{\sigma}_D^2$ and $\hat{\sigma}_E^2$ are not necessarily positive although $\hat{\sigma}_D^2 + \hat{\sigma}_E^2$ is non-negative.

To demonstrate the application of this model, data from a "split-ballot" experiment reported by Kalton, Collins and Brook (1978), were used. For several questions, the sample of 2157 respondents was split into two at random and two design options were applied. Each of the questions is treated separately as a single variable, since the multi-variate data were not available to us. In Table 1 the values of $\hat{\delta}$, the consistent estimator of δ obtained from (6) and (7), are given, as well as the estimated actual design-option effects for (k=2) and the hypothetical design-option effects for (k=6). The results show large variations in response variance, with the highest values for alternatives with respect to the question formulation.

3. Multivariate Multi-factor Extension

For application to more realistic situations we consider the extension of the basic model to the multivariate case with several factors of design-options. Consider the case of two factors A and B with I and J categories, respectively, and assume that all IJ options are tested. Let n_{ij} be the number of observations obtained by option (i,j) and let $Y_{ijk\ell}$ be the observation on the dichotomous variable $\ell(=1,\ldots,L)$, obtained from the k-th observation for the (i,j) option.

Denote:

$\underline{Y}_{ijk} = (Y_{ijk1},\ldots,Y_{ijkL})$, $i=1,\ldots,I$; $j=1,\ldots,J$; $k=1,\ldots,n_{ij}$.

The model assumed is:

$$\underline{Y}_{ijk} = \underline{P} + \underline{A}_i + \underline{B}_j + \underline{C}_{ij} + \underline{E}_{ijk},$$

where $\underline{P}$ is a fixed vector (of length L) and $\underline{A}_i$, $\underline{B}_j$, $\underline{C}_{ij}$ and $\underline{E}_{ijk}$ are uncorrelated (vector) random variables, with

$$E(\underline{A}_i) = E(\underline{B}_j) = E(\underline{C}_{ij}) = E(\underline{E}_{ijk}) = \underline{0}$$

and (unknown) variance-covariance matrices:
$V(\underline{A}_i) = \sum_A$; $V(\underline{B}_i) = \sum_B$; $V(\underline{C}_{ij}) = \sum_C$; $V(\underline{E}_{ijk}) = \sum_E$

Here $\underline{A}_i$ and $\underline{B}_j$ represent main factor effects and $\underline{C}_{ij}$ represents interaction effects.

Let $\hat{\underline{P}}_{ij} = n_{ij}^{-1} \sum_k \underline{Y}_{ijk}$, $\hat{\underline{P}}_{i.} = J^{-1} \sum_j \hat{\underline{P}}_{ij}$, $\hat{\underline{P}}_{.j} = I^{-1} \sum_i \hat{\underline{P}}_{ij}$ and $\hat{\underline{P}} = (IJ)^{-1} \sum_{i,j} \hat{\underline{P}}_{ij}$. Then $\hat{\underline{P}}$ is an unbiased estimator of $\underline{P}$ with variance-covariance matrix:

$$V(\hat{\underline{P}}) = (IJ)^{-1} (\textstyle\sum_A + \sum_B + \sum_C + \bar{n}_H^{-1} \sum_E)$$

where $\bar{n}_H = [(IJ)^{-1} \sum_{ij} 1/n_{ij}]^{-1}$ is the harmonic mean of the sample sizes n_{ij}.

The components of variance-covariance can be estimated unbiasedly by an extension of the "unweighted analysis of means estimators", Searle (1971) to the multivariate situation as follows.

Denote:

$$MSA = J(I-1)^{-1} \sum_i (\hat{\underline{P}}_{i.} - \hat{\underline{P}})(\hat{\underline{P}}_{i.} - \hat{\underline{P}})'$$

$$MSB = I(J-1)^{-1} \sum_j (\hat{\underline{P}}_{.j} - \hat{\underline{P}})(\hat{\underline{P}}_{.j} - \hat{\underline{P}})'$$

$$MSC = [(I-1)(J-1)]^{-1} \{\textstyle\sum_{ij} (\hat{\underline{P}}_{ij} - \hat{\underline{P}}_{i.} - \hat{\underline{P}}_{.j} + \hat{\underline{P}}) (\hat{\underline{P}}_{ij} - \hat{\underline{P}}_{i.} - \hat{\underline{P}}_{.j} + \hat{\underline{P}})'\}$$

$$MSE = (n-IJ)^{-1} \sum_{ijk} (\underline{Y}_{ijk} - \hat{\underline{P}}_{ij})(\underline{Y}_{ijk} - \hat{\underline{P}}_{ij})',$$

where $n = \sum_{ij} n_{ij}$.

Then:

$$\hat{\textstyle\sum}_A = J^{-1} (MSA-MSC)$$

$$\hat{\textstyle\sum}_B = I^{-1} (MSB-MSC)$$

$$\hat{\textstyle\sum}_C = MSC - \bar{n}_H^{-1} MSE$$

$$\hat{\textstyle\sum}_E = MSE$$

are unbiased estimators of the relevant variance-covariance components.

As in the univariate single-factor case, these estimators are not guaranteed to be non-negative definite nor even to have non-negative diagonals (estimates of the variance components).

4. Empirical Example

The above model was applied to part of a pretest for the 1986 National Health Interview Survey conducted by the Bureau of Census for the National Center for Health Statistics in the fall of 1985. The study (CASM, part C) comprised some 400 household interviews and two questionnaire design factors were considered, with respect to a dental health supplement:

A - Questionnaire version - two versions tested: the regular supplement booklet and a special CASM booklet. The differences between the versions related to question formulation (see appendix).

B - Interview type - this factor related to whether the full NHIS core interview was administered prior to the supplement interview or not.

Approximately equal numbers of interviews were allocated to each of the four factor combinations and the components of variance-covariance were estimated according to the methods outlined above for a set of five questions. All the estimates of covariances were very small in absolute value and only results for the variances are reported in the following. In addition to estimates of σ_A^2 and σ_B^2 (the main effect variances) and σ_C^2 (the interaction variance), the total response variance $\sigma_D^2 = \sigma_A^2 + \sigma_B^2 + \sigma_C^2$, and its ratio to the total variance, δ, were estimated - see Table 2.

The fact that in several cases for some of the components and even for the total response error, unbiased non-negative estimators were not obtained, is due to the very small differences between responses for different questionnaire versions or interview types. Even in cases where a positive estimator for the overall response error is obtained, the estimated design-option effects are small.

5. Discussion

Under the very simple assumptions made here, the estimation of the variance components of δ--the proportion of design-option variance out of total variance--allows the investigation of the effect of using different values of k (number of different design-options) in the survey itself on the overall variance. As noted above, for a fixed sample size n, with m=n/k units allocated to each of k design options, the increase in variance (relative to the extreme case k=n) is given by by $1 + (n/k-1)\delta$. This implies that for the usual case in survey practice, where only a single design-option is used (i.e., k=1), even the very small values of δ, as estimated in Tables 1 and 2, result in very large increases in variance--by a factor of $1 + (n-1)\delta$,--especially for large sample sizes.

This raises the possibility of decreasing the design-option variance component by using more than a single design-option in the survey. Obviously, increasing k will increase costs so that a balance between cost and variance has to be found. If good estimates of δ and a realistic cost function are available, the optimal value of k can be found. However, practical limitations will in many cases allow for only a small number of design-options. The use of k>1 gives the important additional benefit of ensuring that an unbiased estimate of design-option variance is made available.

It must be pointed out, however, that the estimates of δ obtained above must be treated with extreme caution, and the variability of these estimates may be considerable. This is brought out by the fact that in many cases the unbiased estimates of variance are negative. Nevertheless, the results give some indication of the large differences in the relative contribution of design-option variance for different types of questions. In the example presented earlier, the order of response categories seems to induce the lowest design-option variances while question formulation and position lead to large variances.

Obviously, a great deal of additional work is required in assessing the effect of choice of design options on response bias and variance. More complex and realistic models need to be developed and the conditions under which the underlying models hold must be verified empirically. The models will also have to relate to results of investigations about the cognitive aspects of different design options. The definition of the relevant universe of design options has to be carefully defined for each case, and the implications of different definitions and the sensitivity of the results to thesevariations and to model alternatives will have tobe assessed. Further extensions to polytomous qualitative variables, multiple choice questions, complex sample designs and complex estimators must also be considered.

This research was partly supported by a National Science Foundation grant (No. SES-8403415).

6. References

Brown, N.R. Rips, L.J. and Shevell, S.K. (1985), "The Subjective Dates of National Events in Very-Long-Term Memory", Cognitive Psychology, 17, 139-177.

Hansen, M.H., Hurwitz, W.N. and Bershad, M.A. (1961), "Measurement Errors in Censuses and Surveys", Bulletin of the International Statistical Institute, 38, 351-374.

Jabine, T.B. and Rothwell, N.D. (1970), "Split Panel Tests of Census and Survey Questionnaires", Proceedings of the Social Statistics Section, American Statistical Association, 4-13.

Kalton, G., M. Collins and L. Brook (1978), "Experiments in Wording Opinion Questions", Appl. Statist. 27, 149-161.

Lessler, J.T. and Sirken, M.G., (1985), "Laboratory-based Research on the Cognitive Aspects of Survey Methodology", Milbank Memorial Fund Quarterly, 63, 565-581.

Lord, F. and Novick R.N. (1968), Statistical Theories of Mental Test Scores, Addison-Wesley, 1968.

Marquis, K.H., Marquis, M.S., and Polich, (1986) "Response Bias and Reliability in Sensitive Topic Surveys", Journal of the American Statistical Association, 81, 381-389.

Monsees, M.L. and Massey, J.T., (1979), "Adapting Procedures for Collecting Demographic Data in a Personal Interview to a Telephone Interview", Proceedings of the Social Statistics Section, American Statistical Association, 130-135.

Searle, S.R. (1971), Linear Models, New York: Wiley.

Sirken, M.G., (1986), "Error Effects of Survey Questionnaires on the Public's Assessment of Health Risks", Am. J. Public Health, 76, 367-368.

Tversky, A. and Kahneman, D., (1974), "Judgments Under Uncertainty: Heuristics and Biases", Science, 185, 1124-1131.

Table 1: Estimates of δ and its Design-Option Effects

Subject of Question	Design Alternatives	$\hat{\delta}$ Estimate of δ	Design option effect $1 + (m-1)\hat{\delta}$ k = 2	k = 6
Priority to buses	With/without reference to cars	.040	43.9	15.3
Bus lanes	With/without preliminary question	.035	39.3	13.7
Lowering of speed limit	Response order			
- to 30 mph		.00002	1.03	1.01
- to 20 mph		.00028	1.30	1.10
- to 40 mph		.00003	1.04	1.01
Lowering of driving standards	Position of question			
- younger drivers		NA	NA	NA
- drivers in general		.0106	12.39	4.79

NA - no positive estimate of response variable available

Table 2: Estimates of Standard Deviation (%) of Variance Components and Proportion of Design-Option Variance

Question	$\hat{\sigma}_A$	$\hat{\sigma}_B$	$\hat{\sigma}_C$	$\hat{\sigma}_D$	$\hat{\sigma}_E$	$\hat{\delta}=\hat{\sigma}_D^2/(\hat{\sigma}_D^2+\hat{\sigma}_E^2)$
Fluoridation of public water supply	0*	0.6	0.9	0*	41.1	0*
Dental visits during past 2 weeks	2.5	4.2	0*	3.2	35.8	0.008
Dental visits during past year	0.8	0.8	0*	0*	43.3	0*
Use of fluoride mouth rinse	3.0	0*	6.2	4.6	41.0	0.012
Use of fluoride tablets/ supplements	3.2	3.5	0*	3.0	35.8	0.007

* Unbiased estimator is negative

APPENDIX

Comparison of Regular and CASM Questions on Dental Health

Item	Regular	CASM Questions
Public Water Fluori-dation	Is your home drinking water supply part of a PUBLIC water system, or is it from a well, spring, or cistern? Is YOUR home drinking water supply FLUORIDATED?	Does the water that you drink at home come from a public water system or is it from another source, such as a well? Has this public water supply had fluoride added to it?
2-Week Dental Visits	During the two weeks (outlined in red on that calendar), beginning Monday (date) and ending this past Sunday (date), did anyone in the family go to a dentist? Include all types of dentists, such as orthodontists, oral surgeons, and all other dental specialists, as well as dental hygienists. Who was this? During those 2 weeks, did anyone else in the family go to a dentist? During those 2 weeks, how many times did -- go to a dentist?	Is there a particular dentist's office, dental clinic, or some other place that -- usually goes for dental care? Altogether, how many DIFFERENT PLACES do family members go for dental care? Does anyone in the family go to an orthodontist? Who is this? Anyone else? When -- needs to go to the dentist, who usually makes the appointment for --? When -- needs to go to the dentist, how does -- usually get there? The item above was followed by the Regular version of 2-Week Dental Visits question.
12-month Dental Visits	During the past 12 months {that is, since (12-month date) a year ago}, how many visits did -- make to a dentist? (Include the (Number) visit(s) you already told me about.)	To help you remember possible visits I will read a list of reasons some people have for going to the dentist. Some people go to the dentist for a check-up and to have their teeth cleaned or to have a tooth filled or capped. Some people go because they are in pain or because a tooth broke or a filling fell out. Some people go as part of a series of treatments for gum disease, a root canal, or to have a false tooth fitted. And some go as part of a series of orthodontic treatments - to have their teeth straightened. The item above was followed by the Regular version of 12-month Dental Visits question.
Fluoride Mouth-rinse	Recently, some MOUTH RINSES have been developed that contain FLUORIDE to reduce tooth decay. Does anyone in the family now use a mouth rinse that contains FLUORIDE, such as ACT, Fluorigard, Listermint with StanCare, or a similar product? Who is this? Anyone else? Does -- use this fluoride mouth rinse at home, at school, or at work?	Sometimes people use fluoride to protect their teeth. For example, some mouth rinses contain fluoride, others do not. Schools and work places may have fluoride mouth rinse programs. Does anyone in the family take part in such a program? Who is this? Is this at school or at work? Is anyone else in a mouth rinse program? Sometimes fluoride mouth rinses are used at home. Does anyone in the family use a mouth rinse or mouth wash that has fluoride in it? Who is this? Anyone else? What is the name of the fluoride mouth rinse that -- uses?
Fluoride Products	Does -- now use FLUORIDE tablets, drops, or FLUORIDE vitamin supplements which are intended to be SWALLOWED?	Sometimes doctors or dentists prescribe pills or drops with fluoride in them. Does anyone in the family now take vitamins with fluoride in them? Who is this? Anyone else? Does anyone in the family now take any other kind of fluoride drops, pills or tablets? Who is this? Anyone else?