MODELING THE EFFECTS OF QUESTION ORDER AND FORM ON SURVEY RESPONSES

James H. Drew, GTE Laboratories Incorporated
40 Sylvan Rd., Waltham, MA 02254

KEY WORDS: questionnaire effects, structural equations, measurement errors

Background

It is virtually an article of faith among applied survey researchers that questionnaire effects exist and play a substantial role in determining survey results. The context, order, wording and response format of a question, among many other things, are all felt to affect the response, but there have been few attempts to formalize such effects. Some of the substantive findings in the literature have been summarized by Sudman and Bradburn (1974), Schuman and Presser (1981) and Groves (1989). Andrews (1982,1984) used a structural equation version of Fiske and Campbell's' multitrait-multimethod analysis to model the effects of various item, response and interview modes for a large number of substantive question topics.

The work we discuss below uses a structural equations/measurement error model to confirm the importance of questionnaire effectsafter close relations due to substantive relationships have been explicitly modeled. In much of the structural equations literature (see Joreskog 1990, for example), unmodeled survey item relationships are accounted for by unrestricted covariances. In this paper, we propose slightly more structure by testing directionality among the item relationships,.testing the effect of item spacing in the questionnaire, and comparing the size of the effects of item spacing, response category similarity, and differences in respondent task.

Similarity of the subject matter of a particular item is also an obvious source of association among item responses, but we shall use a structural model to account (at least partially) for this general effect.

Questionnaire Effects in a Survey of General Telephone Service

We analyze a telephone survey of local telephone customers in which their recent local service, in its various facets, is evaluated. an overall quality question at the beginning of the questionnaire (OVQ_b) is immediately followed by a local dial quality item (LOCQ) and a series of items about network problems, here summarized as a dichotomous variable PROBEX-problem existence). Shortly thereafter is a question on long distance quality (LDQ), and after another brief set of items, a question on billing quality (BILLQ). The final item in the survey is a virtual repeat of the opening overall quality item (OVQ_e).

In the model we introduce here, we take explicit account of possible differences between survey items, such as OVQ_b and OVQ_e, and the theoretical concepts they putatively measure. Since we know from a previous study (Drew and Bolton, 1991) that the two items differ dramatically in their responses, it is prudent to initially model each as an indicator of its own corresponding construct, labeled *initial quality* and *final quality*. Using the arguments of the consumer behavior literature (e.g. Helson, 1964), the final quality construct is a function of initial quality, performance and disconfirmation. The attribute measures of local dial, long distance and billing (LOCQ, LDQ and BILLQ, respectively) are indicators of both current performance and (because they too are attitudes) disconfirmation, and therefore may be products of both overall quality constructs.

The prime reason for the difference in attitude ratings between the first and last overall quality measure appears to be the intervention of items with specific reference periods. For instance, a the survey asks how the respondent would rate local dial service *over the last 30 days*. Neither overall quality measure has a stated reference period but there is some evidence that a reference period effect operates for the last overall quality item, but not the first. See Drew and Bolton (1991) for more information. The existence of static and network problems in the reference period (PROBEX) is a disconfirmation measure and is thus a product of the *final quality* construct.

A preliminary fit of the model just described gives a large chi-square value, indicating a gross lack of fit to the observed correlation matrix for these variables. The theoretical considerations of the preceding paragraphs, however, suggest some structure which militates against the wholesale postulation of unconstrained covariances among the observed survey items. More preliminary models, for example, show the insignificance of any direct covariances between PROBEX and any other survey items, and between LDQ and BILLQ. We have argued that responses to an item depend on the item's position in the questionnaire, and now postulate that each measure is somehow related to the major attribute measure just preceding it in the questionnaire. This phenomenon has been called reactivity in the social science literature (Sullivan and Feldman, 1979). In our survey, OVQ_b is related to LOCQ, which in turn is related to LDQ, and BILLQ is related to OVQ_e. Furthermore, similarity of items, or of response categories may generate a relationship, while dissimilarity dampens one. Thus OVQ_b is related to OVQ_e, while LOCQ and PROBEX are not directly related despite their proximity in the questionnaire.

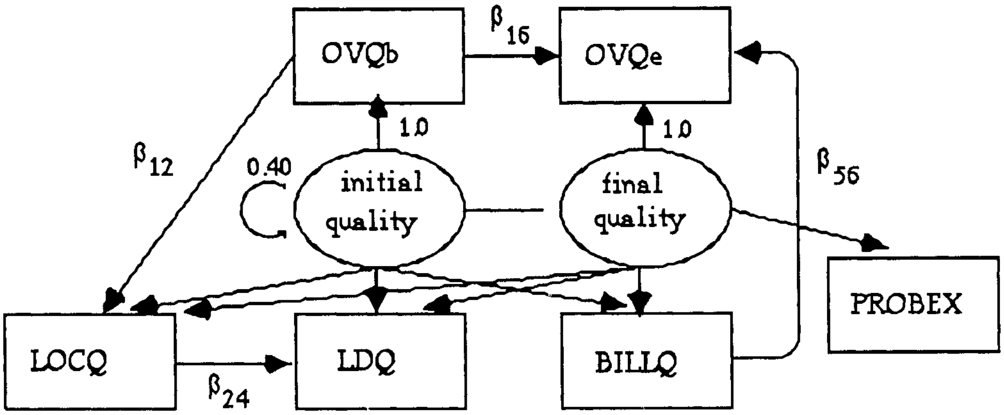
A next step up in structure is to postulate unconstrained covariances between these items. Allowing the pairs (OVQ_b, LOCQ), (OVQ_b, OVQ_e), (LOCQ, LDQ). and (BILLQ, OVQ_e) to co-vary independently of the mutual dependence on the two quality constructs results in a model whose chi-square value is 5.57 with 5 d.f.

The relative sizes of the covariance estimates is of interest, since these measure the strength of the inter-item relationships when an underlying structure has been established. The results suggest an interesting interplay between the effects of item spacing and cognitive task similarity. The largest covariance estimate is between LOCQ and LDQ, presumably because these items are both closely spaced and require evaluations of similar services (one a major subset of the other). The covariance between the fairly closely spaced, but cognitively somewhat different items LOCQ and OVQ_b is nearly the same size as the covariance between the virtually equivalent but far-spaced OVQ_b and OVQ_e. The covariance between the fairly closely-spaced but cognitively different items LDQ and BILLQ is indistinguishable from 0. Finally and most complex is the large negative covariance among BILLQ and OVQ_e; the significance suggests a memory effect between the closely-spaced items, but the sign suggests the effect of a difference in cognitive processing characterized by billing being a minor component of overall quality. The values are displayed in the table below. Except for Cov(LDQ, BILLQ), all the values are significantly different from zero at $\alpha=0.05$.

Pair	Covariance Estimate
OVQ_b,LOCQ	0.1458
LOCQ,LDQ	0.2872
LDQ, BILLQ	0.0600
BILLQ, OVQ_e	-0.2368
OVQ_b, OVQ_e	0.1940

The non-significance of the chi-square value from the unrestricted covariance model leads us to consider other constraints which are in greater agreement with previous informal conjectures on survey item effects. Now we attempt to impose directional structure on these relationships. Note that this is not a more constrained model, since each covariance parameter of the preceding model is replaced by a corresponding regression coefficient. Numbering the items OVQ_b, LOCQ, PROBEX, LDQ, BILLQ, and OVQ_e according to their relative position in the questionnaire and letting β_{ij} represent the regression coefficient for the effect of item i on item j, the null model we consider can be diagrammed as in Figure I.

Figure 1



To fit this model, some special handling of the data and its correlation coefficients is needed. Since each survey response is ordinal, polychoric correlations should be used instead of the usual product-moment variety, and some merging of categories is necessary so that these correlations are consistent with the postulate of underlying multivariate normality. To ensure the identification of the labeled coefficients, the coefficients of relations between the two quality latent variables and their indicators are both assumed to be 1.0, and the variance of the initial quality latent variable is estimated to be 0.40, based on previous repeated measurement studies. Since there are some items missing for some respondents, the effective sample size in the model fitting is taken to be the smallest pairwise sample size among all the values in the correlation matrix; this is of course a conservative choice. This model will be loosely called the "unconstrained model."

The fit of this model seems to be acceptable. The chi-square value indicating the closeness of its predicted correlation matrix with the observed polychoric correlation matrix is an insignificant 9.12 with six degrees of freedom. More important is the fact that the residuals from the fit of the polychoric correlation matrix shows no systematic patterns, has elements with values unimodally distributed about zero, and that the largest value of the standardized residuals is only 1.72.

We wish to formally test two hypotheses:

1) that each β_{ij} labeled in the diagram is non-zero, and

2) whether the two latent quality variables are distinct.

The chi-square values associated with the unconstrained model, with the four models with one of the β_{ij}'s set equal to 0, and with the model equating the latent variables, are given below.

Model	d.f.	Chi-Square
Unconstrained	6	9.12
$\beta_{12}=0$	7	24.22
$\beta_{24}=0$	7	44.03
$\beta_{56}=0$	7	41.83
$\beta_{16}=0$	7	93.66

It follows from these values that each β_{ij} appears to be non-zero. For instance, the test that $\beta_{12} = 0$ consists of comparing 24.22 - 9.12 = 15.10 to a chi-square random variable with 7 - 6 = 1 degree of freedom. This value is significant at the 0.0001 level. Likewise, the three other tests yield highly significant test statistics. We conclude that proximate item relationships exist apart from their being common indicators of latent variables.

To add evidence that the relationships among the indicators are products of their questionnaire position, we consider the reversal of some of the regression coefficients, in contrast to their relative questionnaire placement. One might conceive of OVQ_b being an outcome of OVQ_e, since the latter may be a more extensive or well-considered measure. Unfortunately, the model is not identified if a reciprocal relationship is first tested, against which the two alternative causal directions are compared. However, we can examine the fit of the model given in the figure above, with β_{16} replaced by β_{61} so OVQ_b is a function of OVQ_e. Fitting this new model leads to residuals with a reasonably normal appearance, but with a chi-square value of 20.68 on five degrees of freedom ($p<0.001$). The significance of this statistic, along with the non-significance of the model pictured above, casts considerable doubt on a relation between these two variables which is in opposition to their survey placement.

Similarly, it is not inconceivable that OVQ_e helps determine the response to BILLQ, rather than the opposite order which is consistent with questionnaire position. However, this model's chi-square is a significant 12.54 with five degrees of freedom ($p<0.03$), and so we reject the "reversal" of the regression coefficient. Reversing the other relations between the indicators can be tested, but the models make little substantive sense since is is unlikely that long distance quality (LDQ) affects local dial quality (LOCQ) or that LOCQ affects OVQ_b.

It should be noted that when this structural model is correct, some of its results differ dramatically from those obtained by ordinary least squares performed on the indicator variables. For instance, a basic analysis of overall quality should include calculating the effect of service attributes on the true quality evaluation. A modification of the structural model given above can be made to allow this estimation. In addition to the overall quality factors, and the relationships between proximate or similar indicators, one can postulate three other factors corresponding to local dial, long distance and billing which generate the attribute indicators. The attribute factors together determine the overall quality factor through a linear structural model. With the data from this experiment, and letting the four factors be denoted by f_{ovq}, f_{locq}, f_{ldq}, and f_{billq}, respectively, the structural equation portion of the model is estimated as

$$f_{ovq} = 0.745\, f_{locq} + 0.594\, f_{ldq} - 0.329\, f_{billq}\,.$$

Only the first coefficient is significant at the 0.05 level. In contrast, a standard OLS regression of the indicator OVQ_b on LOCQ, LDQ and BILLQ yields

$$OVQ_b = 0.521\ LOCQ - 0.114\ LDQ + 0.312\ BILLQ,$$

where in the both cases, the indicators were first standardized, so that no intercept term is necessary in the last equation. Each coefficient is highly significant. The considerable difference in the size and ordering of the two sets of coefficients is largely due to the other terms in the preceding model, particularly the relations between the designated survey items.

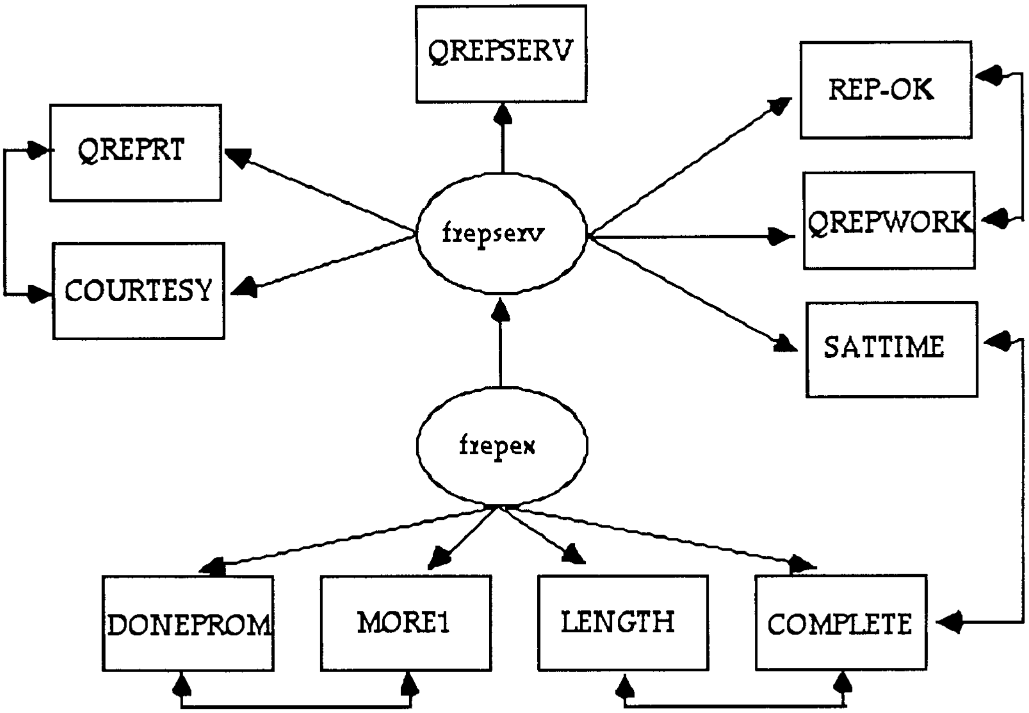
Questionnaire Effects in a Survey of Telephone Repair Incidents

We present below a structural equation model of 1225 responses to a survey of customers who had reported and experienced a repair of their local telephone service. Following an overall evaluation of the entire repair experience (QREPSERV), customers respond to series of items about their report of the problem (QREPRT, the general evaluation of the reporting process; and COURTESY, the courtesy of the repair representative), their evaluation of the repair work (QREPWRK, an overall evaluation;

SATTIME, satisfaction with the time taken for the work, and REPOK, the acceptability of the work), and some objective characteristics of the experience (MORE1, whether more than one repair report needed; DONEPROM, whether the work was done when promised; LENGTH, the total length of the repair incident; and COMPLETE, whether the customer was notified when the repair was complete).

Building a model for responses to this questionnaire is of substantial interest because these items encompass both factual and evaluative questions, their response categories and the item wordings themselves have several forms, and the items are variously spaced in the questionnaire. Without considering questionnaire effects, the model represented by Figure 2 was constructed as a base for the study of questionnaire effects.

Figure 2



In this path diagram, covariances have been postulated between items whose subject matter might suggest a correlation based on more than a common latent variable. The two such variables are *frepex*, summarizing the factual execution of the repair, and *frepserv*, its subjective evaluation. Preliminary tests showed no need for a further splitting of the evaluation variable into a report evaluation and a repair work evaluation. The fit of this model to the observed polychoric correlation matrix is very poor, being $X^2 = 1191.87$ with 30 degrees of freedom.

The fit of this basic model can be improved dramatically by adding paths between the following pairs of items, with the justification for the inclusion of each being given in the following table:

Variable Pair	Reason
QREPRT, REPOK	Item Wording- "Rate . . . Quality . . ."
COURTESY, SATTIME	Similar Response Categories- "Very . . ., Somewhat . . ."
COURTESY, QREPWORK	Item Proximity
QREPWORK, MORE1	Item Proximity
REPOK, COMPLETE	Item Proximity

When these relationships are added to the model, with the earlier item considered as an input to a later item, the resulting chi-square is a substantially improved $X^2 = 323.10$ with 23 degrees of freedom. The magnitude of this reduction in chi-square value is strong evidence for the existence of these effects. Furthermore, the coefficients associated with these paths are significant, and comparable in size to the path coefficients relating item responses and their latent variables. It follows, then, that these relationships play an important role in determining the relative impact of service characteristics and attribute evaluations on overall evaluations.

Conclusions

We have developed structural equation/measurement error models for sets of items from single surveys. Contrary to the widespread practice of relating these items or their underlying constructs based on their intrinsic meanings, our models demonstrate that, in this context at least, relations among the items must also be based on such survey characteristics as the items' relative questionnaire positions and their response categories. When these relations are integrated into the models of perceived quality developed for two telephone company surveys, the quality attributes and their coefficients are found to differ in many significant ways from simple models not accounting for measurement error and item interrelationships. These models provide a somewhat more quantitative form for these effects than is commonly seen in the survey and sociology literature, and lays some groundwork for the study of the generalizability of these effects.

References

Andrews, F. (1984). Construct Validity and Error Components of Survey Measures: A Structural Modelling Approach, *Public Opinion Quarterly,* Summer 1984, 48,2, 409-422.

Drew, J. and R. Bolton (1991). The Structure of Customer Satisfaction: Effects of Survey Measurement, to appear in *The Journal of Customer Satisfaction, Dissatisfaction and Complaining Behavior.*

Groves, R.M. (1989). *Survey Errors and Survey Costs.* New York: Wiley-Interscience.

Helson, H. (1964). *Adaptation-Level Theory.* New York: Harper and Row Publishers.

Joreskog, K., A.-M. Aish and I. Munck (1990). *Using LISREL for Analyzing Measurement Errors in Surveys.* Unpublished Manuscript

Schuman, H. and S. Presser (1981). *Questions and Answers in Attitude Surveys.* New York: Academic Press.

Sudman, S. and N. Bradburn (1974). *Response Effects in Surveys.* Chicago: Aldine.

Sullivan, J., and S. Feldman (1979). *Multiple Indicators: An Introduction.* Sage University Paper series on Quantitative Applications in the Social Sciences. Beverly Hills and London: Sage Publications.