

Adjusting for mode of administration effect in
surveys using mailed questionnaire and
telephone interview data

Karl Bang Christensen

National Institute of Occupational Health, Denmark

Helene Feveille

National Institute of Occupational Health, Denmark

Svend Kreiner

Department of Biostatistics, University of Copenhagen

Jakob Bue Bjorner

QualityMetric, Inc.

July 3, 2006

Abstract

Psychometric scales and ordinal categorical items are often used in surveys with different modes of administration like mailed questionnaires and telephone interviews. If items are perceived differently by people over the telephone this can be a source of bias. We describe how such bias can be tested using item response theory and propose a method to correct for such differences using multiple imputation and item response theory. The method is motivated and illustrated by analyzing data from an occupational health study using mailed questionnaire and telephone interview data to evaluate job influence.

Keywords: Item response theory, survey, multiple imputation, mode of administration, telephone interview.

Introduction

Many surveys use different modes of administration, e.g. mail, telephone, or in-person administration, and the effect that the mode of administration has on responses can affect the results (McHorney, Kosinski, & Ware, 1994). While mixed modes of administration can be used to improve response rates, differences in response behavior can lead to biased results.

In a study comparing respondents to mailed questionnaires to non respondents subsequently interviewed by telephone (Brambilla & McKinlay, 1987) socioeconomic differences were disclosed. Controlling for these differences did not, however, remove differences in reported health outcomes between mail and telephone respondents.

When responses are not comparable across modes of administration the situation can be regarded as a missing data problem in the senses that, e.g. persons responding by mailed questionnaire have a missing telephone interview response. Multiple imputation (Rubin, 1976, 1987) replaces each missing value with a number, say five, plausible values that represent the uncertainty about the right value to impute. Each of the resulting complete data sets are analyzed, and the results are then combined generating valid statistical inferences that properly reflect the uncertainty due to missing values.

In a recent study Powers and colleagues estimated the differences in self-rated health by mode of administration and used multiple imputation to make self-rated health comparable for telephone and mail administered questionnaires (Powers, Mishra, & Young, 2005). The effect of mode of administration on changes in mental health scores were of a magnitude that is considered to be clinically meaningful. Multiple imputation yielded effect estimates that were similar for telephone and mail respondents.

Multiple imputation is thus an attractive method that may be used to adjust scale scores for mode of administration bias. For the purpose of surveys, imputing data on the sum score scale is of interest, because underlying latent variables are seldom used in reporting of results. This paper describes how translation based on the joint distribution of sum scores can be done using item response theory models. Doing this yields a framework where differences in response behavior can be tested, interpreted, and taken into account.

Statistical model

In this section the model is described in the situation where subjects are randomized to either mailed questionnaire or telephone interview. Persons who are randomized to one mode of administration, but for some reason respond using another method are excluded, as are those with incomplete item responses. Translation from observed responses to mailed questionnaires to telephone interview responses is used as an example.

Let Y_i denote the response to item i , for $i = 1, \dots, I$, and let X denote an indicator taking the value one if the person has responded by telephone interview and zero otherwise. The probability of each of the ordinal categorical item responses $h = 1, \dots, m_i$ are modeled using a generalized partial credit model (Muraki, 1992) where item parameters depend on mode of administration

$$P(Y_i = h | \theta, X = x) = \frac{\exp(\sum_{v=1}^h \alpha_i^{(x)} (\theta - \beta_{iv}^{(x)}))}{\sum_{c=1}^{m_i} \exp(\sum_{v=1}^c \alpha_i^{(x)} (\theta - \beta_{iv}^{(x)}))} \quad (1)$$

Where $(\alpha_i^{(0)}, \beta_{i1}^{(0)}, \dots, \beta_{im_i}^{(0)})$ are item parameters for the questionnaire sample, $(\alpha_i^{(1)}, \beta_{i1}^{(1)}, \dots, \beta_{im_i}^{(1)})$ are item parameters for the telephone sample, and θ is the person parameter. The α 's are discrimination parameters and the β 's are threshold parameters (the locations on the θ -axis where there is equal probability of adjacent response categories).

Assuming that item responses are conditionally independent given the person parameter (local independence), the probability of a response vector $y = (y_1, \dots, y_I)$, is given by the product

$$P(Y_1 = y_1, \dots, Y_I = y_I | \theta, X = x) = \prod_{i=1}^I P(Y_i = y_i | \theta, X = x) \quad (2)$$

The model is identified if $\beta_{i1}^{(x)} = 0$ for $x = 0, 1$ and all i and the distribution of the latent variable θ in the population is assumed to be standard normal. This implies that there are no differences between the mailed questionnaire sample and the telephone interview sample with respect to the trait being measured - an assumption that is realistic when subjects are randomized to either mailed questionnaire or telephone interview. The marginal distribution is then

$$P(Y_1 = y_1, \dots, Y_I = y_I | X = x) = \int \prod_{i=1}^I P(Y_i = y_i | \theta, X = x) \varphi(\theta) d\theta \quad (3)$$

where $\varphi(\theta) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}\theta^2)$. A likelihood approach based on these probabilities will yield marginal maximum likelihood estimates of item parameters, and a likelihood ratio test of the null hypothesis

$$H_0 : (\alpha_i^{(0)}, \beta_{i1}^{(0)}, \dots, \beta_{im_i}^{(0)}) = (\alpha_i^{(1)}, \beta_{i1}^{(1)}, \dots, \beta_{im_i}^{(1)}) \quad (4)$$

can be computed. Other hypotheses could also be of interest, e.g. equal

thresholds but different discriminations. For the sum score $S = \sum_{i=1}^I Y_i$, the distribution conditional on the mode of administration X can be computed by summation

$$P(S = s|\theta, X = x) = \sum_{(y_1, \dots, y_I)}^{(*)} P(Y_1 = y_1, \dots, Y_I = y_I|\theta, X = x) \quad (5)$$

over the set over response vectors (y_1, \dots, y_I) with $\sum_{i=1}^I y_i = s$. If a person were to respond to both set of items independently the joint probabilities would be $P(S = s_1|\theta, X = 1)P(S = s_0|\theta, X = 0)$ and this can be used to compute the joint marginal probabilities

$$Q(s_1, s_0) = \int P(S = s_1|\theta, X = 1)P(S = s_0|\theta, X = 0)\varphi(\theta)d\theta. \quad (6)$$

Based on these joint marginal probabilities the conditional distribution of telephone responses given questionnaire responses

$$Q(s_1|s_0) = \frac{Q(s_1, s_0)}{\sum_s Q(s, s_0)} \quad (7)$$

can be computed.

Translation of scores between modes of administration

A model based on the probabilities, Equation 1, enables translation of observed sum scores from mailed questionnaires to telephone interview sum scores. This can be done using the joint probabilities, Equation 6. Given a mailed questionnaire scale score s_0 , that is observing $(S, X) = (s_0, 0)$, an estimate of the sum score on the telephone interview scale can be given in several ways, e.g. as the conditional mean $\sum_s sQ(s|s_0)$ or as the most likely telephone interview sum score: that is s_1 , for which $Q(s_1|s_0) = \max_s Q(s|s_0)$. All that is required is computation of the joint probabilities, Equation 6.

These are reasonable translations for individual responses, but the uncertainty is ignored. When the focus is on groups rather than single persons it is preferable to use multiple imputation (Rubin, 1976, 1987) that replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute. This has been implemented in version 9 of SAS.

Based on item response theory and plausible values, the approach is as follows: (i) upon estimation of item parameters, the conditional probabilities, Equation 7, are computed, (ii) for the persons in the sample who have responded by questionnaire, and thus have a missing telephone interview score, a number, say five, of plausible telephone interview score are drawn from the conditional distribution, and (iii) the imputed data sets are analyzed and results of are combined generating valid statistical inferences that properly reflect the uncertainty due to missing values.

Example

In 1997 the Danish National Institute of Occupational Health conducted a study of the psychosocial work environment in a general population sample of people between 20 and 60 years (Kristensen, Hannerz, Høgh, & Borg, 2005). Persons were randomized for mailed questionnaires (2/3 of the sample) or for telephone interviews (1/3 of the sample). Data from 1609 employees are considered here.

Here responses to four items about influence are considered: 'Do you have a large degree of influence concerning your work?', 'Do you have a say in choosing who you work with?', 'Can you influence the amount of work assigned to you?', and 'Do you have any influence on WHAT you do?' (response options: Always, Often, Sometimes, Seldom, Never/hardly ever).

When disregarding persons with incomplete item responses the simple sum score, for these four items with five response categories, has 17 categories. This sum score is often transformed linearly to a zero to 100 scale when reporting results.

For the questionnaire sample the mean score was 54.9 and the standard deviation was 23.1 For the telephone interview sample the mean score was 55.8 and the standard deviation was 25.8. Using the folded F statistic ($\max(s_1^2, s_0^2) / \min(s_1^2, s_0^2)$, where s_1^2 and s_0^2 are the sample variances in the two samples) the variances were found to be significantly different ($p = 0.0017$). The means did not differ significantly.

The generalized partial credit model, Equation 1, was fitted to these four items. Inspection of observed and expected mean item scores showed an excellent fit to the model for the mailed questionnaire sample, and an acceptable fit for the telephone interview sample (results not shown). The estimated item parameters are shown in Table 1. For all items the thresholds

are closer together for the telephone interview sample, and discriminations are lower. This indicates that extreme categories are used more often in the telephone interview sample. A possible explanation for this is difficulties in remembering and distinguishing between five response categories over the telephone. This would also explain why the fit of the model was worse for the telephone interview sample. The fact that the extreme categories dominate also has the consequence that thresholds are also disordered. An example of item characteristic curves are shown in Figure 1.

This difference in response behavior explains why the variance is larger in the telephone interview sample. The likelihood ratio test of the hypothesis, Equation 4, of equal item parameters across mode of administration is highly significant ($\chi^2(20) = 180.5, p = 0.000$). This means that the telephone interview and questionnaire samples should not uncritically be collapsed.

Based on these item parameters the probabilities, Equation 6, in the joint distribution of sum scores were computed. An illustration is shown in Figure 2, where darker shades of grey indicate higher probability.

Two of the main purposes for collecting data like these are surveillance, e.g. comparison of job groups, and epidemiology, e.g. evaluation of the association between work environment and sickness absence. Table 2 shows the difference between job groups - means, standard deviations and pairwise T tests in the two samples. The same trend is seen in both samples, but the differences are only significant in the telephone sample. This is not due to differences in sample size. It is clear from this that results can depend crucially on mode of administration.

Table 3 shows the results of Poisson regression analysis of the effect of influence on number of sickness absence days. For the telephone interview sample the result is borderline significant. Imputation of scores on the tele-

phone interview sample scale for the 1012 persons with a missing telephone interview score (the mailed questionnaire sample) resulted in data sets with $597+1012=1609$ subjects and the combined results of the analyses of these data sets, yielded a significant effect of the same magnitude as the small sample analysis.

Discussion

Although a mixed mode approach may reduce non response bias, more research is required concerning the reasons for response differences between modes and to eliminate any differences caused by problems in data quality. Multiple imputation methods have been proposed as a solution to this problem (Powers et al., 2005). Here multiple imputation using an item response theory model was discussed and implemented.

Item response theory models are useful for equating between scales using information from persons who have responded to both set of items. This paper described a situation where no persons have responded using both modes of administration, and the population was used as the link between the two measurement scales. The assumption that the θ 's are independent and identically distributed, i.e. that the true level of the persons are independent of the method used, is crucial. This assumption is reasonable because of the randomization. The normality assumption could be relaxed.

The method of equating sum scores is quite general in the sense, that other item response theory models, like the graded response model can also be used. The approach can also be applied to single ordinal items.

Multiple imputation methods have been used in item response theory in sampling designs that provide information about population characteristics, but administer too few items to estimate the latent trait individually (Mislevy, 1991; Mislevy, Beaton, Kaplan, & Sheehan, 1992). In these applications marginal estimation procedures are approximated by constructing plausible values of individual values of the latent trait.

The example showed that results can depend critically on mode of administration, and that using multiple imputation to include information can increase the statistical power of the analyses. Using item response theory

models the differences in response behavior could be tested, and interpreted.

The multiple imputation based method described and implemented here is mainly of interest when the purpose is statistical analysis and not the study of individual levels. If the focus is on individual, e.g. in a clinical setting or in educational testing, the translation based on the joint distribution of sum scores, Equation 6, can still be used, but the conditional mean is likely to be a better prediction than imputation. In this case the illustration of the joint probabilities in Figure 2 can yield information about the specificity of the prediction.

Importantly, the method is only feasible for complete case analysis. One solution in the presence of incomplete item responses is to impute item by item, this could however cause the variability to be underestimated because the uncertainty in the item parameter estimates is ignored. Further research on the performance and properties of this method is thus warranted.

References

- Brambilla, D. J., & McKinlay, S. M. (1987). A comparison of responses to mailed questionnaires and telephone interviews in a mixed mode health survey. *American Journal of Epidemiology*, *126*(5), 962–971.
- Kristensen, T. S., Hannerz, H., Høgh, A., & Borg, V. (2005). The copenhagen psychosocial questionnaire - a tool for the assessment and improvement of the psychosocial work environment. *Scandinavian Journal of Work, Environment & Health*, *31*, 438–449.
- McHorney, C. A., Kosinski, M., & Ware, J. E. (1994). Comparisons of the costs and quality of norms for the sf-36 health survey collected by mail versus telephone interview: results from a national survey. *Medical Care*, *32*, 551–567.
- Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177–196.
- Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*, 133–161.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176.
- Powers, J. R., Mishra, G., & Young, A. F. (2005). Differences in mail and telephone responses to self-rated health: use of multiple imputation in correcting for response bias. *Australian and New Zealand Journal of Public Health*, *29*, 149–154.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Figure 1: An example of item characteristic curves: the item 'Do you have a large degree of influence on decisions about your work?'

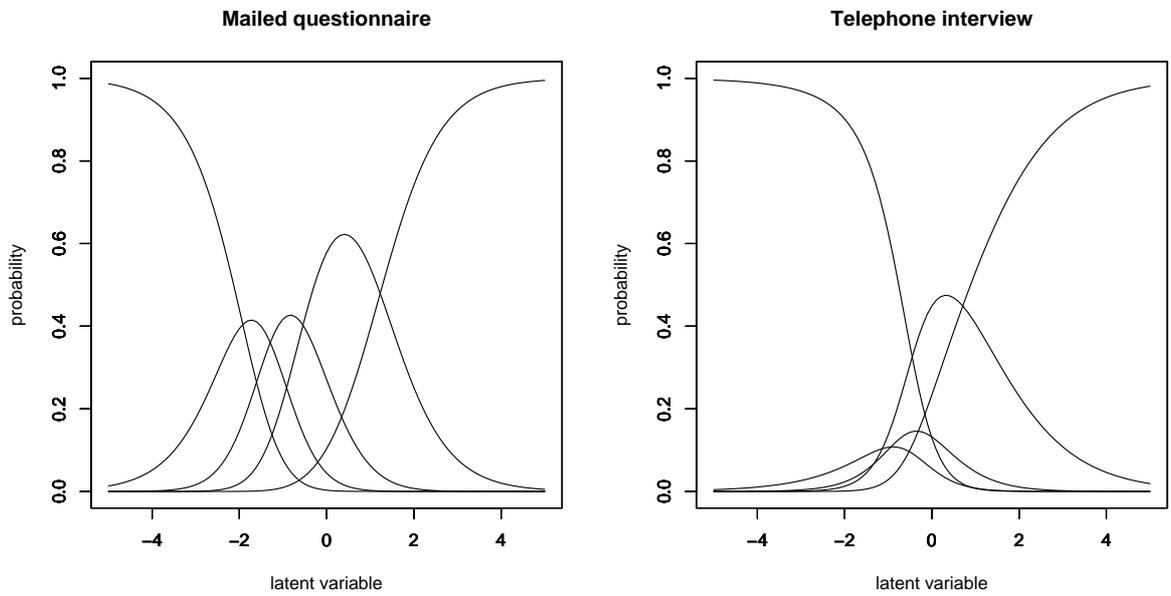


Table 1: Estimated parameters for the four items about influence

Item	Mailed Questionnaire		Telephone interview	
	Par.	Est. (S.E.)	Par.	Est. (S.E.)
... decisions	$\alpha_i^{(0)}$	1.40 (0.13)	$\alpha_i^{(1)}$	0.92 (0.12)
	$\beta_{i2}^{(0)}$	-1.92 (0.14)	$\beta_{i2}^{(1)}$	-0.94 (0.23)
	$\beta_{i3}^{(0)}$	-1.29 (0.10)	$\beta_{i3}^{(1)}$	-1.33 (0.21)
	$\beta_{i4}^{(0)}$	-0.54 (0.08)	$\beta_{i4}^{(1)}$	-1.03 (0.18)
	$\beta_{i5}^{(0)}$	1.23 (0.09)	$\beta_{i5}^{(1)}$	0.71 (0.13)
... who you work with	$\alpha_i^{(0)}$	0.83 (0.08)	$\alpha_i^{(1)}$	0.71 (0.10)
	$\beta_{i2}^{(0)}$	-0.51 (0.12)	$\beta_{i2}^{(1)}$	0.49 (0.24)
	$\beta_{i3}^{(0)}$	-0.02 (0.12)	$\beta_{i3}^{(1)}$	-0.10 (0.21)
	$\beta_{i4}^{(0)}$	0.38 (0.12)	$\beta_{i4}^{(1)}$	0.30 (0.21)
	$\beta_{i5}^{(0)}$	1.55 (0.15)	$\beta_{i5}^{(1)}$	0.68 (0.21)
... amount of work	$\alpha_i^{(0)}$	0.54 (0.05)	$\alpha_i^{(1)}$	0.53 (0.07)
	$\beta_{i2}^{(0)}$	-1.13 (0.19)	$\beta_{i2}^{(1)}$	0.25 (0.29)
	$\beta_{i3}^{(0)}$	-0.19 (0.17)	$\beta_{i3}^{(1)}$	-0.34 (0.28)
	$\beta_{i4}^{(0)}$	0.40 (0.17)	$\beta_{i4}^{(1)}$	-0.31 (0.26)
	$\beta_{i5}^{(0)}$	1.27 (0.20)	$\beta_{i5}^{(1)}$	0.45 (0.24)
... what you do	$\alpha_i^{(0)}$	1.35 (0.14)	$\alpha_i^{(1)}$	1.04 (0.16)
	$\beta_{i2}^{(0)}$	-1.45 (0.11)	$\beta_{i2}^{(1)}$	-0.18 (0.24)
	$\beta_{i3}^{(0)}$	-0.85 (0.09)	$\beta_{i3}^{(1)}$	-1.37 (0.22)
	$\beta_{i4}^{(0)}$	-0.40 (0.08)	$\beta_{i4}^{(1)}$	-0.59 (0.15)
	$\beta_{i5}^{(0)}$	0.66 (0.07)	$\beta_{i5}^{(1)}$	0.39 (0.12)

Figure 2: Illustration of the probabilities in the joint sum score distribution (darker shades of grey indicate higher probability).

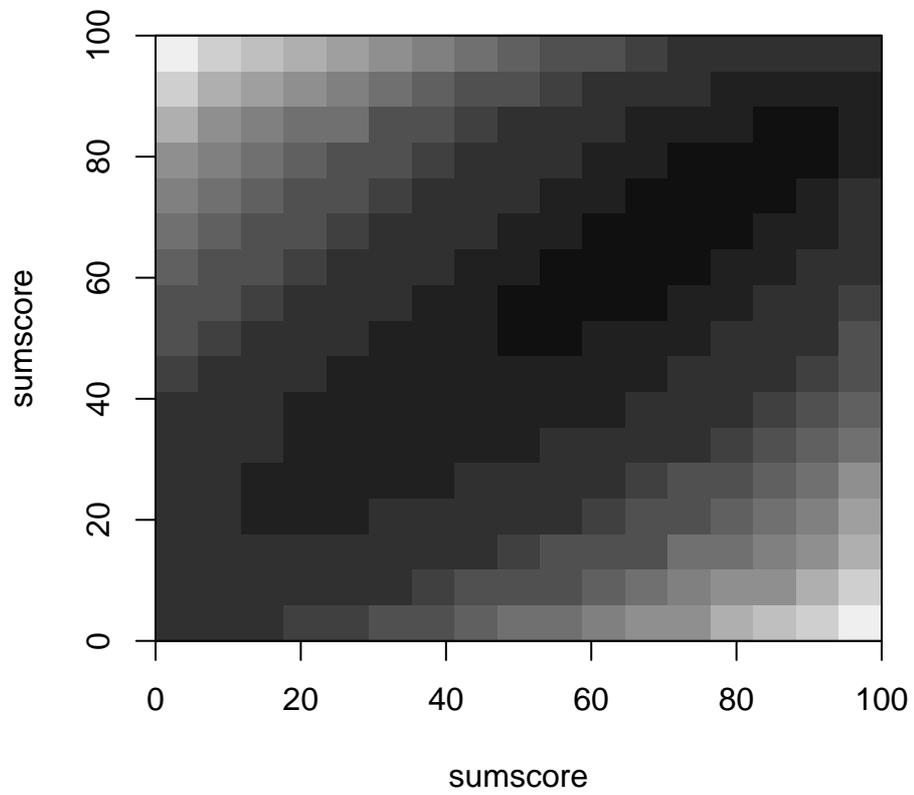


Table 2: The difference between job groups - means, standard deviations and pairwise T tests in the two samples.

	Telephone sample			Questionnaire sample		
	N	Mean (S.D.)	T test	N	Mean (S.D.)	T test
Trade workers	44	62.5 (25.3)	-	51	54.7 (21.3)	-
Industry workers	62	48.1 (24.2)	p=0.0037	71	50.1 (21.6)	p=0.2487
Office workers	82	44.0 (26.4)	p=0.0002	170	47.3 (21.4)	p=0.0324

Table 3: The effect of influence on sickness absence: Poisson regression controlled for gender and age. Multiple imputation imputes scores on the telephone sample scale for the questionnaire sample.

	N	RR*	95% CI
Influence - telephone sample	597	0.96	(0.92, 1.00)
Influence - multiple imputation	1609	0.96	(0.93, 0.98)

*: Rate ratios show the effect of a ten point increase in influence.