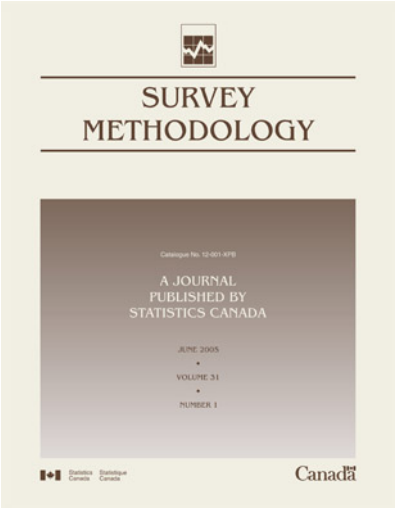
Catalogue no. 12-001-XIE

Survey Methodology

December 2005



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

December 2005

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2006

All rights reserved. The content of this electronic publication may be reproduced, in whole or in part, and by any means, without further permission from Statistics Canada, subject to the following conditions: that it be done solely for the purposes of private study, research, criticism, review or newspaper summary, and/or for non-commercial purposes; and that Statistics Canada be fully acknowledged as follows: Source (or "Adapted from", if appropriate): Statistics Canada, year of publication, name of product, catalogue number, volume and issue numbers, reference period and page(s). Otherwise, no part of this publication may be reproduced, stored in a retrieval system or transmitted in any form, by any means—electronic, mechanical or photocopy—or for any purposes without prior written permission of Licensing Services, Client Services Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

May 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 31, No. 2, pp. 169-182
Statistics Canada, Catalogue No. 12-001

Variance-Covariance Functions for Domain Means of Ordinal Survey Items

Alistair James O'Malley and Alan Mark Zaslavsky [1]

Abstract

Estimates of a sampling variance-covariance matrix are required in many statistical analyses, particularly for multilevel analysis. In univariate problems, functions relating the variance to the mean have been used to obtain variance estimates, pooling information across units or variables. We present variance and correlation functions for multivariate means of ordinal survey items, both for complete data and for data with structured non-response. Methods are also developed for assessing model fit, and for computing composite estimators that combine direct and model-based predictions. Survey data from the Consumer Assessments of Health Plans Study (CAHPS®) illustrate the application of the methodology.

Key Words: Variance function; Correlation function; Hierarchical model; Ordinal response; Nonresponse; Skip pattern.

1. Introduction

Survey data are often used to obtain measures for comparisons across estimation domains. In our motivating example, surveys are conducted to elicit reports on experiences with health plans (entities administering health care) from enrolled members; similarly a survey might assess schools by administering tests to a sample of students.

An essential part of the analysis of survey data is the calculation of sampling variances, or the sampling-covariance matrix of a multivariate estimator. The standard survey sampling approach is to compute variances directly for each estimator in each domain. Direct variance estimates may be unstable when the number of respondents to an item is small because the sample size for a domain is small, because the item is applicable to only a fraction of respondents (such as users of specialized equipment in health surveys), or because we are interested in means for a small subgroup (such as those with chronic illnesses).

By modeling variance estimates as functions of the unit (domain) means, we can pool information across units to obtain more stable estimates. Although modeling may introduce bias, for small units this is offset by the reduction in sampling variation. One may also consider generalizing variance estimates across items in addition to or instead of domains. This will be appropriate when there are groups of items for which the same mean-variance relationship is likely to hold. However, when there are many more domains than items, the greatest potential gain is from generalizing across domains rather than across items.

A *Generalized Variance Function* (GVF) is a mathematical model describing the relationship between the variance or relative variance of a survey estimator and its expectation. When multiple estimates are produced from the same sample, Wolter (1985, chapter 5) proposes the model

$$V / M^2 = \theta_0 + \theta_1 / M,$$

where M and V denote the expected value and variance of the estimator respectively. Such a form might be suitable for variables such as income or wealth for which a nearly constant coefficient of variation might be plausible because the mean and standard deviation are proportional to the length of the reference period. Modeling the coefficient of variation is thus most suited to situations where the variables are similar in content but have different scales with unrestricted ranges (*e.g.*, income collected monthly and yearly). In our problem the items are ordinal and so a model of the coefficient of variation is not a natural choice. Other proposed GVFs also have simple forms (Woodruff 1992; Otto and Bell 1995).

If a suitable GVF can be found, it can simplify calculations and make variance estimates more stable. Furthermore, summarizing sampling variance estimates in the form of a function also facilitates presentation of large volumes of statistics (Wolter 1985, pages 201-202). Finally, modeling variances as functions of means facilitates iterative re-estimation of sampling variances in hierarchical modeling. In practice the decision to use variance functions in a hierarchical modeling context depends on the goodness of the fit of the GVF; only with a sufficiently good fit is use of the GVF worthwhile.

Past work on GVFs is relatively sparse. Wolter (1985, chapter 5) gave an overview but provided only a few references, as did Valliant, Dorfman and Royall (2000,

1. Alistair James O'Malley and Alan Mark Zaslavsky, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115-5899, U.S.A. E-mail: omalley@hcp.med.harvard.edu and zaslavsk@hcp.med.harvard.edu.

pages 344–348). Valliant (1992a, 1992b) used GVFs to smooth time-dependent indices in time series analysis. Woodruff (1992) used GVFs for variance estimation of employment change in the Current Employment Survey, and Wolter (1985, pages 208–217) illustrates the use of GVFs on data from the Current Population Survey. GVFs are also used in the National Health Interview Survey (Valliant *et al.* 2000, page 344).

Huff, Eltinge, and Gershunskaya (2002) and Cho, Eltinge, Gershunskaya and Huff (2002) considered GVFs for the United States Current Employment Survey and Consumer Expenditure Survey. Eltinge (2002) uses GVFs to estimate a full sampling covariance matrix when samples are too small to produce stable estimates for all areas, estimating the components of the mean squared error (MSE) of the GVF model. Otto and Bell (1995) fit GVFs to median income, per capita income, and age-group poverty rates in the Current Population Survey, assuming an autoregressive dependence between rates over time and a Wishart distribution for the sampling covariance matrices.

Our research extends previous research on GVFs in four directions. First, we use the GVF to generalize across domains rather than items. Thus, we do not assume that different items have the same GVF, although it might be reasonable to fit models of the same form for items with similar response categories. Second, we develop GVFs for the full covariance matrix, which must be estimated for joint inference on multiple outcomes. Thirdly, we focus on the relationship between means and variances of items with the ordinal response formats often used in survey questionnaires, rather than on homoscedastic continuous responses. Finally, we explicitly allow for patterns of nonresponse due to structured skip patterns. While structured item nonresponse can be ignored (except for its effect on sample size) in univariate estimation, it must be considered explicitly to model bivariate relationships because it affects the sampling covariance of item means. Furthermore, because the number of responses varies across items, we cannot model the sampling covariances using a Wishart distribution, which has only a single parameter for sample size.

We first describe direct estimation of variances and covariances, including the case when data are missing due to skip patterns. In section 3 we introduce models for generalized variance and covariance functions (GVCFs) and lay out our strategies for model fitting and evaluation and for combining direct estimates and model predictions. In section 4, we apply our methods to a major health care survey. In section 5, we conclude by describing applications and extensions of our methods.

2. Direct Estimates of Sampling Variances of Domain Means

We index observations by domain h, items (indices i and j), and respondents (indices k and l); $y_{h,ik}$ and $r_{h,ik}$ denote the outcome and response indicator of subject k in domain h on item i. We suppress the index for item when referring to all items for a respondent or domain, and have no need for the subscript for respondent when discussing the means, variances, and correlations of items.

Direct estimation of the sampling covariance matrix of domain means (henceforth, "variance estimation") begins by expressing the means as functions of totals of the outcomes and response indicators. We replace $y_{h,ik}$ with 0 for missing observations so that totals are defined in the presence of skip patterns. Following the notation of Särndal, Swenson and Wretman (1992, pages 24–28; 36–42), let U_h and S_h describe the population and sample respectively for the h^{th} domain, $Y_{h,i} = \sum_{U_h} y_{h,ik}$, $R_{h,i} = \sum_{U_h} r_{h,ik}$, $\hat{Y}_{h,i} = \sum_{S_h} \breve{y}_{h,ik}$, and $\hat{R}_{h,i} = \sum_{S_h} \breve{r}_{h,ik}$, where $\breve{y}_{h,ik} = y_{h,ik} / \pi_{h,k}$, $\breve{r}_{h,ik} = r_{h,ik} / \pi_{h,k}$, and $\pi_{h,k} = \text{pr}(k \in S_h)$.

The vector of mean outcomes for the population of elements within domain h is

$$M_h = f(Y_h, R_h) = \left(\frac{Y_{h,1}}{R_{h,1}}, \ldots, \frac{Y_{h,I}}{R_{h,I}} \right),$$

where $Y_h = (Y_{h,1}, \ldots, Y_{h,I})$ and $R_h = (R_{h,1}, \ldots, R_{h,I})$. An estimator is

$$f(\hat{Y}_h, \hat{R}_h) = \left(\frac{\hat{Y}_{h,1}}{\hat{R}_{h,1}}, \ldots, \frac{\hat{Y}_{h,I}}{\hat{R}_{h,I}} \right).$$

A first order Taylor series expansion of $f(\hat{Y}_h, \hat{R}_h)$ about $f(Y_h, R_h)$ produces the approximation

$$\text{var}(f(\hat{Y}_h, \hat{R}_h)) \approx V_h = f'(Y_h, R_h)\, \text{var}(\hat{Y}_h, \hat{R}_h) f'(Y_h, R_h)^T,$$

where $f'(Y_h, R_h)$ is the Jacobian of $f(Y_h, R_h)$. Often it is computationally easier to first calculate $u_{h,k} = f'(Y_h, R_h) z_{h,k}$, where $z_{h,k} = (y_{h,k}, r_{h,k})$, and then evaluate the variance as

$$\begin{aligned} V_h &= \text{var}\left(\sum_{S_h} \breve{u}_{h,k} \right) \\ &= \text{var}\left(\sum_{U_h} \breve{u}_{h,k} I_{h,k} \right) \\ &= \sum_{k,l \in U_h} \Delta_{h,kl} \breve{u}_{h,k} \breve{u}_{h,l}^T, \end{aligned}$$

where $I_{h,k} = 1$ if $k \in S_h$ (indicating that the k^{th} member of domain h is sampled) and 0 otherwise, $\Delta_{h,kl} = \pi_{h,kl} - \pi_{h,k}\pi_{h,l}$, and $\pi_{h,kl} = \text{pr}(k, l \in S_h)$. An estimator for V_h is

$$\hat{V}_h = \sum_{k,l \in S_h} \breve{\Delta}_{h,kl}\, \breve{u}_{h,k}\, \breve{u}_{h,l}^T, \qquad (1)$$

where $\breve{\Delta}_{h,kl} = \Delta_{h,kl} / \pi_{h,kl}$.

To describe evaluation of $\hat{V}_h$ one need only consider one diagonal (*i.e.*, variance) element and one off-diagonal (*i.e.*, covariance) element. The sub-matrix of the Jacobian formed by the i^{th} and j^{th} items is given by

$$f'(Y_h, R_h) = \begin{pmatrix} \dfrac{1}{R_{h,i}} & 0 & -\dfrac{Y_{h,i}}{R_{h,i}^2} & 0 \\ 0 & \dfrac{1}{R_{h,j}} & 0 & -\dfrac{Y_{h,j}}{R_{h,j}^2} \end{pmatrix}.$$

For, $z_{h,k} = (y_{h,ik}, y_{h,jk}, r_{h,ik}, r_{h,jk})$, it follows that

$$u_{h,k} = f'(Y_h, R_h)\, z_{h,k} = \begin{pmatrix} \dfrac{1}{R_{h,i}}(y_{h,ik} - M_{h,i}\, r_{h,ik}) \\ \dfrac{1}{R_{h,j}}(y_{h,jk} - M_{h,j}\, r_{h,jk}) \end{pmatrix},$$

where $M_{h,i} = Y_{h,i} / R_{h,i}$ is the mean outcome of the i^{th} item in domain *h*. Hence,

$$\hat{V}_{h,ii} = \frac{1}{R_{h,i}^2} \sum_{k,l \in S_h} \breve{\Delta}_{h,kl} (\breve{y}_{h,ik} - M_{h,i}\, \breve{r}_{h,ik})(\breve{y}_{h,il} - M_{h,i}\, \breve{r}_{h,il}) \quad (2)$$

and

$$\hat{V}_{h,ij} = \frac{1}{R_{h,i}\, R_{h,j}} \sum_{k,l \in S_h} \breve{\Delta}_{h,kl} (\breve{y}_{h,ik} - M_{h,i}\, \breve{r}_{h,ik}) \times (\breve{y}_{h,jl} - M_{h,j}\, \breve{r}_{h,jl}). \qquad (3)$$

To evaluate (2) and (3), we make a further approximation by substituting $\hat{R}_{h,i} = \sum_{S_h} \breve{r}_{h,ik}$ and $\hat{M}_{h,i} = \sum_{S_h} \breve{y}_{h,ik} / (\sum_{S_h} \breve{r}_{h,ik})$ for $R_{h,i}$ and $M_{h,i}$.

When sampling rates are small, or if we wish to make predictions for a large super-population (*e.g.*, all potential enrollees in a health plan, not just those currently enrolled), $\breve{\Delta}_{h,kl} = 1 - \pi_{h,k} \approx 1$ if $k = l$, $\breve{\Delta}_{h,kl} \approx 0$ if $k \neq 1$, and the sampling design approaches sampling with replacement. Under the sampling with replacement design, approximately unbiased estimators are

$$\hat{V}_{h,ii} = \frac{1}{\hat{R}_{h,i}^2} \sum_{k \in S_h} (\breve{y}_{h,ik} - \hat{M}_{h,i}\, \breve{r}_{h,ik})^2 \qquad (4)$$

and

$$\hat{V}_{h,ij} = \frac{1}{\hat{R}_{h,i}\, \hat{R}_{h,j}} \sum_{k \in S_h} (\breve{y}_{h,ik} - \hat{M}_{h,i}\, \breve{r}_{h,ik})(\breve{y}_{h,jk} - \hat{M}_{h,j}\, \breve{r}_{h,jk}). \qquad (5)$$

These estimators can be generalized to accommodate clustering.

With equal-probability sampling within domains, (4) and (5) reduce to

$$\hat{V}_{h,ii} = \frac{1}{\hat{R}_{S_{h,i}}^2} \sum_{k \in S_h} (y_{h,ik} - \hat{M}_{h,i}\, r_{h,ik})^2 \qquad (6)$$

and

$$\hat{V}_{h,ij} = \frac{1}{\hat{R}_{S_{h,i}} \hat{R}_{S_{h,j}}} \sum_{k \in S_h} (y_{h,ik} - \hat{M}_{h,i}\, r_{h,ik})(y_{h,jk} - \hat{M}_{h,j}\, r_{h,jk}), \qquad (7)$$

where $\hat{R}_{S_{h,i}}$ is the number of respondents to item *i* in domain *h*.

3. Models for Variance Functions

In this section we propose specifications for models for variances and for sample correlations with complete responses or with structured skipped responses. We then discuss model fitting and evaluation strategies. We assume that these domains are nonoverlapping strata, so the sampling errors for different domains are independent.

We transform the ordinal ratings to the $[0,1]$ interval by the transformation $p_{h,i} = (B_{h,i} - M_{h,i}) / (B_{h,i} - A_{h,i})$, where $A_{h,i}$ and $B_{h,i}$ are the minimum and maximum response categories for item *i* in domain *h* respectively. We focus on modeling variances for large values of $M_{h,i}$ (small values of $p_{h,i}$) because in our motivating example mean outcomes are typically near the high end of the scale.

3.1 Variance Functions

To account for the variable number of respondents over domains and items, and differing scales, we normalize the variance estimators in (6) for sample size and re-scale:

$$\tilde{V}_{h,ii} = \frac{\hat{R}_{S_{h,i}}\, \hat{V}_{h,ii}}{(B_{h,i} - A_{h,i})^2}.$$

With unequal probability sampling within domains, a normalization factor could be used that accounts for the weights. One possible normalization is to multiply $\hat{V}_{h,ii}$ by $\hat{R}^*_{S_{h,i}} = (\sum \breve{r}_{h,ik})^2 / (\sum \breve{r}^2_{h,ik})$, where $\breve{r}_{h,ik}$ is the response indicator for item *i* for the k^{th} subject in the h^{th} domain, in place of $\hat{R}_{S_{h,i}}$. This approximation, proposed in Kish (1965), has a model based justification (Gabler, Haeder and Lahiri 1999). It works well if the sampling probabilities vary modestly in the sample, but can lead to inefficiency if the variation is excessive (Korn and Graubard 1999, page 173; Spencer 2000).

Because the items in our example have ordinal scales, the variance must go to 0 as $p_{h,i} \to 0$ or $p_{h,i} \to 1$. An obvious predictor with this property is the variance function of the Bernoulli distribution, $p_{h,i}(1 - p_{h,i})$. This holds exactly for

dichotomous items, and might be a useful approximation for items with three or more categories.

As alternatives to the Bernoulli variance model we considered models with a variety of polynomial and other functions of the means as predictors. Of all the models considered, the quadratic family of models were found to fit as well as any. We focused on the following quadratic models.

$$\text{Model V1:}\quad \tilde{V}_{h,ii} = \beta_{1i} p_{h,i}, \tag{8}$$

$$\text{Model V2:}\quad \tilde{V}_{h,ii} = \beta_{2i} p_{h,i}(1-p_{h,i}), \tag{9}$$

$$\text{Model V3:}\quad \tilde{V}_{h,ii} = \beta_{1i} p_{h,i} + \beta_{2i} p_{h,i}(1-p_{h,i}). \tag{10}$$

Thus we consider a linear variance model V1, a binomial-like model V2, and a general quadratic variance model V3. All models correctly ensure $\tilde{V}_{h,ii} = 0$ when $p_{h,i} = 0$, but only V2 ensures that $\tilde{V}_{h,ii} = 0$ when $p_{h,i} = 1$. The rationale behind V1 is that relationships are often approximately linear over small intervals. Both V1 and V2 are submodels of the two-parameter quadratic V3. We also considered models for $\log(\tilde{V}_{h,ii})$, but these models did not fit as well.

The model V3 is equivalent to the model suggested by Wolter (1985, chapter 5); the equivalence is seen by expressing the right-hand side of V3 in terms of $p_{h,i}$ and $p_{h,i}^2$, and then dividing both sides by $p_{h,i}^2$ to obtain the relative variance. However, parameter estimates obtained by fitting the two forms of the model may be different depending on the modeling assumptions used.

3.2 Correlation Functions with Complete Data

Because correlations are independent of the scale of the data, we model the correlations and derive the sampling covariances, rather than modeling the covariances directly. We model the sample correlations

$$\hat{\rho}_{h,ij} = \frac{\hat{V}_{h,ij}}{(\hat{V}_{h,ii}\,\hat{V}_{h,jj})^{1/2}},$$

via the unrestricted transformed values $Z_{h,ij} = \log\{(1+\hat{\rho}_{h,ij})/(1-\hat{\rho}_{h,ij})\}$. Unlike the variance models, models for correlations may include an unrestricted intercept, since there is no natural restriction on the correlation when $p_{h,i}$ or $p_{h,j}$ approaches 0 or 1.

Because $\hat{\rho}_{h,ij}$ is a function of the first and second moments of items *i* and *j*, it seemed reasonable to first focus on linear and quadratic models for $Z_{h,ij}$. As with variance functions, we found that a more extensive range of models (*e.g.*, models with logarithms of the means as predictors) did not substantially improve model fit. We ultimately focused on the following nested series of models.

$$\text{Model C1:}\quad Z_{h,ij} = \alpha_{0ij}, \tag{11}$$

$$\text{Model C2:}\quad Z_{h,ij} = \alpha_{0ij} + \alpha_{3ij}\, p_{h,i}\, p_{h,j}, \tag{12}$$

$$\text{Model C3:}\quad Z_{h,ij} = \alpha_{0ij} + \alpha_{1ij}(p_{h,i} + p_{h,j}) + \alpha_{3ij}\, p_{h,i}\, p_{h,j}, \tag{13}$$

$$\text{Model C4:}\quad Z_{h,ij} = \alpha_{0ij} + \alpha_{1ij} p_{h,i} + \alpha_{2ij} p_{h,j} + \alpha_{3ij}\, p_{h,i}\, p_{h,j}, \tag{14}$$

$$\text{Model C5:}\quad Z_{h,ij} = \alpha_{0ij} + \alpha_{1ij} p_{h,i} + \alpha_{2ij} p_{h,j} + \alpha_{3ij}\, p_{h,i}\, p_{h,j} + \alpha_{4ij}\, p_{h,i}^2 + \alpha_{5ij}\, p_{h,j}^2, \tag{15}$$

Model C3 is model C4 with the constraint $\alpha_{1ij} = \alpha_{2ij}$.

3.3 Predicting Covariances with Structured Missing Data

When the data have skip patterns, the sample correlations of the ratings for the set of respondents who answered both items can be modeled by (11)–(15), as in the complete response case. The corresponding sample covariances can be easily estimated by using the fitted variance functions to re-scale the predicted correlations. However, because the sampling covariance reflects the variability in the whole sampling process, not just the variability within the sub-population of respondents who answered both items, the relationship between sample covariance and sampling covariance is more complicated than if the data were complete. In this section we derive the relationship between the sample covariance for the set of respondents who answered both items and the sampling covariance. This allows correlation models such as (11)–(15) to be applied to data with skip patterns.

There are four distinct data patterns for any pair of items: response to both items, one response and one skipped item (two patterns), and both items skipped. We extend our notation by introducing a superscript representing the response status of a second item. Let $\hat{Y}_{h,ij}^1 = \sum_{S_h} \breve{y}_{h,ik}\breve{r}_{h,jk}$, $\hat{Y}_{h,ij}^0 = \sum_{S_h} \breve{y}_{h,ik}(1-\breve{r}_{h,jk})$, $\hat{R}_{h,ij}^1 = \sum_{S_h} \breve{r}_{h,ik}\breve{r}_{h,jk}$, $\hat{R}_{h,ij}^0 = \sum_{S_h} \breve{r}_{h,ik}(1-\breve{r}_{h,jk})$, $\hat{M}_{h,ij}^1 = \hat{Y}_{h,ij}^1 / \hat{R}_{h,ij}^1$, $\hat{M}_{h,ij}^0 = \hat{Y}_{h,ij}^0 / \hat{R}_{h,ij}^0$. Then

$$\hat{M}_{h,i} = \frac{\hat{R}_{h,ij}^1 \hat{M}_{h,ij}^1 + \hat{R}_{h,ij}^0 \hat{M}_{h,ij}^0}{\hat{R}_{h,i}}.$$

In the equal probability sampling case, substitution of the above expression for $\hat{M}_{h,i}$ into (7) yields

$$\tilde{V}_{h,ij} = \frac{\hat{R}_{h,ij}^1}{\hat{R}_{h,i}\,\hat{R}_{h,j}} \left\{ \hat{C}_{h,ij}^1 + \frac{\hat{R}_{h,ij}^0 \hat{D}_{h,ij} \hat{R}_{h,ji}^0 \hat{D}_{h,ji}}{\hat{R}_{h,i}\,\hat{R}_{h,j}} \right\}, \tag{16}$$

where $\hat{D}_{h,ij} = \hat{M}^1_{h,ij} - \hat{M}^0_{h,ij}$. Here $\hat{C}^1_{h,ij} = \sum_S (\breve{y}_{h,ik} - \hat{M}^1_{h,ij}\,\breve{r}_{h,ik})(\breve{y}_{h,ik} - \hat{M}^1_{h,ji}\,\breve{r}_{h,jk}) / \hat{R}^1_{h,ij}$ is the normalized sample covariance of the ratings for the set of respondents who answered both items (which can be predicted using correlation and variance functions, and in the case of unequal probability sampling applying a normalization factor). When the sampling probabilities are not equal, Equation (16) holds exactly only if $\sum_S \breve{r}_{h,jk}(\breve{y}_{h,ik} - \hat{M}^1_{h,ik}\,\breve{r}_{h,ik}) = 0$. Therefore, (16) may be expected to provide a good approximation if the sampling probabilities for one item are not highly correlated with the residuals for another item. In general, the appropriateness of using (16) for unequal probability sampling designs should be checked.

The estimated mean differences $\hat{D}_{h,ij}$ determine the contribution of the response pattern to the sampling covariance. Either $\hat{D}_{h,ij}$ or $\hat{D}_{h,ij}\,\hat{D}_{h,ji}$ may be modeled in the process of obtaining smoothed estimates of $\tilde{V}_{h,ij}$. In our application, the $\hat{D}_{h,ij}$ were typically small. Because the second term of (16) is a product of two factors of small magnitude ($\hat{D}_{h,ij}$ and $\hat{D}_{h,ji}$), the contribution of $\hat{D}_{h,ij}$ to (16) was small and it sufficed to use a simple model for $\hat{D}_{h,ij}$, such as a constant for each item pair. However, unique constants should be estimated for each pair of items.

3.4 Model Fitting and Evaluation

We estimate the parameters of the variance or correlation function using iteratively reweighted least squares regression. Weighting is important when the number of responses varies greatly across domains, as in our motivating example.

In this section we index domains (h) and respondents (k) but not items as the same methodology applies to each variance and correlation model. Exact computations are derived for the equal probability sampling case, and approximations are noted for the unequal probability sampling case. Generically, the direct estimators $\tilde{f}_h$, true values f_h, and model predictions $\hat{f}_h$ are related through the hierarchical model

$$\text{Level I}: \quad \tilde{f}_h = f_h + \in_h, \tag{17}$$

$$\text{Level II}: \quad f_h = \hat{f}_h + e_h, \tag{18}$$

where $\in_h \sim [0, \sigma^2_h / \hat{R}_{S_h}]$, $e_h \sim [0, \tau^2]$, and $[\mu, \sigma^2]$ indicates a distribution with expectation μ and variance σ^2 but unspecified form. In the unequal probability sampling case we replace $\hat{R}_{S_h}$ with $\hat{R}^*_{S_h}$. Here $\in_h$ represents sampling error and e_h represents model error. Marginally, $\tilde{f}_h = \hat{f}_h + e_h + \in_h$ so in the regression we weight the observation for domain h by $w_h = (\tau^2 + \sigma^2_h / \hat{R}_{S_h})^{-1}$, the inverse of the marginal variance. With equal-probability sampling, the variance of the direct estimate of $\sigma^2_h = E[\tilde{f}_h - f_h]^2$ is given by

$$\hat{\sigma}^2_h(\tilde{f}_h) = \frac{1}{\hat{R}_{S_h} - 1}\left\{\frac{1}{\hat{R}_{S_h}}\sum_{k\in S}(y_{h,k} - \hat{M}_h r_{h,k})^4 - (1 - \frac{3}{\hat{R}_{S_h}})\tilde{f}^2_h\right\}$$
if f is a variance (19)

$$\hat{\sigma}^2_h(\tilde{f}_h) = \frac{4}{\hat{R}_{S_h} - 3} \text{ if } f \text{ is a transformed correlation.} \tag{20}$$

In the equal probability sampling case Equation (19) is exact and does not depend on parametric assumptions (Seber 1977, page 14). The asymptotic approximation (20) to the variance of the transformed correlation Z_h (Freund and Walpole 1987, page 477) deteriorates as sample sizes decrease, and fails altogether for $\hat{R}_{S_h} \le 3$. However, domains with small sample sizes have little impact on the fitted models; we exclude domains with $\hat{R}_{S_h} \le 3$ from correlation modeling.

When the sampling probabilities are not equal, the large sample counterpart to (19), given by

$$\hat{\sigma}^2_h(\tilde{f}_h) = \sum_{k\in S}\left\{\begin{array}{l}\dfrac{(\breve{y}_{h,k} - \hat{M}_h\breve{r}_{h,k})^2}{\sum_{l\in S}\breve{r}^2_{h,l}} - \dfrac{2w_h}{\sum_{l\in S}\breve{r}_{h,l}} \\ \times(\breve{y}_{h,k} - \hat{M}_h\breve{r}_{h,k}) - \dfrac{\tilde{f}^2_h}{\sum_{l\in S}\breve{r}^2_{h,l}}\breve{r}^2_{h,k}\end{array}\right\}^2,$$

where $w_h = (\sum_S \breve{y}_{h,l}\breve{r}_{h,l}) / \sum_S \breve{r}^2_{h,l}) - \hat{M}_h$, may be used. In the equal probability sampling case, $w_h = 0$ and the above expression reduces to a non-bias corrected version of (19). If the sampling probabilities are not equal, we suggest replacing (20) with the design-effect-corrected estimator

$$\hat{\sigma}^2_h(\tilde{f}_h) = \frac{4}{\hat{R}^*_{S_h} - 3}.$$

The model error variance τ^2 is estimated as:

$$\hat{\tau}^2 = \max\left\{0, \hat{\text{MSE}} - \frac{\sum \hat{R}_{S_h}\hat{\sigma}^2_h(\tilde{f}_h)}{\sum \hat{R}_{S_h}}\right\},$$

where $\hat{\text{MSE}} = \sum_h q_h(\tilde{V}_h - \hat{f}_h)^2$, $q_h = N\hat{R}_{S_h} / \sum_h \hat{R}_{S_h}$, and $N = \sum_h I(\hat{R}_{S_h} > 0)$. The weights are then re-estimated as $\hat{w}_h = (\tau^2 + \hat{\sigma}^2_h(\tilde{f}_h) / \hat{R}_{S_h})^{-1}$, and the GVCF models are refit, iterating to convergence. We again suggest replacing $\hat{R}_{S_h}$ with $\hat{R}^*_{S_h}$ if the sampling probabilities are not equal.

We compared the predictive accuracy of models using $R^2 = 1 - \hat{\text{MSE}} / \hat{\text{MSV}}$, where $\hat{\text{MSE}}$ is the mean squared error of the regression, and $\hat{\text{MSV}}$ is the sample size weighted average of the sampling variances of the direct estimators (variances or transformed correlations) for each

domain. Note that we could have $R^2 < 0$ for a very poorly fitting model.

3.5 Combined Estimators

For domains with small samples, direct survey variance estimates often are too imprecise to be useful, while estimates for larger domains in the same study may be quite reliable. Fay and Herriot (1979) and Ghosh and Rao (1994) demonstrated that shrinking direct estimates towards a model-based smoothed value can lead to substantial gains in precision. They proposed composite or empirical Bayes estimators that are weighted averages of direct and model-based estimators. That is, instead of either using the direct estimates or estimates obtained from generalized variance/ covariance modeling, we use a weighted average of the two estimators to potentially obtain even better estimates.

Such weighted estimators can be constructed for domain variances using the model specified in (17) and (18). A natural approach is to weight the direct model-based estimators inversely proportional to the corresponding sampling and model error variances respectively (denoted σ_h^2 and τ^2 respectively for domain h). The resulting estimator for domain h (for variances and transformed correlations) is:

$$\tilde{f}_h = \frac{\hat{\tau}^2 \tilde{f}_h^{\text{dir}} + \hat{\sigma}_h^2 \tilde{f}_h^{\text{mod}}}{\hat{\tau}^2 + \hat{\sigma}_h^2} = \tilde{f}_h^{\text{dir}} + \frac{\hat{\sigma}_h^2}{\hat{\tau}^2 + \hat{\sigma}_h^2}\left(\tilde{f}_h^{\text{mod}} - \tilde{f}_h^{\text{dir}}\right),$$

where $\tilde{f}_h^{\text{dir}}$ and $\tilde{f}_h^{\text{mod}}$ denote the direct and model-based estimators. This generic formula applies to the variance estimates for all items, and correlation estimates for all pairs of items. The right-most expression has the form of an empirical Bayes estimator.

If the direct and model-based variance estimators are independent, the variance of the resulting combined estimator is $\tau^2\sigma_h^2/(\tau^2 + \sigma_h^2) \le \min\{\tau^2, \sigma_h^2\}$. Thus the composite is as least as precise as either of its two component estimators, improving on ad hoc selection between direct and model-based predictions. This is a useful strategy especially when model-based predictions improve on direct estimates for some, but not all domains.

4. Example: CAHPS® Data Set

The Consumer Assessments of Health Plans Study (CAHPS®) survey (Goldstein, Cleary, Langwell, Zaslavsky and Heller 2001) was designed primarily to elicit consumer ratings and reports on health plans. Plan mean scores (perhaps after recoding) on the various survey items are calculated and reported to consumers, health plans, and purchasers. Each analytic domain consists of the enrollees of a health plan (or geographically defined portion of one) in a year; most of the plans are sampled in multiple years. The stratum is the reporting unit (plan or portion thereof) in a given year; reporting units corresponded to plans with the exception of a few large plans that had multiple reporting units. Therefore, there are many units for variance and covariance function estimation.

We illustrate our methods with a CAHPS data set for beneficiaries of U.S. Medicare managed care plans, a system of private but government-funded entities serving from 5.7 to 6.9 million elderly or disabled beneficiaries in each year during our study period (1997 to 2001). Our data represent 381 reporting domains each sampled in 1 to 5 years for a total of 932 distinct reporting unit by year domains with 705,848 responses. Because samples are drawn independently each year, patients may be sampled in multiple years. However, repeated sampling is rare and can be overlooked for our analysis. Therefore, the domains are strata with equal probability element sampling performed within each. Note that in CAHPS analyses no corrections are made for finite-population sampling since the data are collected to guide choices for future years rather than to record experiences of the specific population in a particular year.

CAHPS items use a variety of ordinal response formats with either 11, 4, 3, or 2 response options. Overall ratings of doctor, specialist, care, and plan are measured on a 0 to 10 scale from "worst possible" to "best possible". Other items use a 4–point ordinal "frequency" scale (never/sometimes/ usually/always), or a 3–point ordinal "problem" scale (not a problem/somewhat a problem/a big problem), or are dichotomous (no/yes). Many items are answered only by respondents who used particular services or had particular needs, as determined by screener items. For example, an item about whether advice was obtained successfully by telephone is only answered by those who first reported that they attempted to obtain advice in that way.

4.1 Descriptive Statistics

Table 1 presents response distributions and domain mean distributions by item type. Missing observations due to structured skip patterns often occurred in blocks, with as many as 11 items skipped on the basis of a single screening question. Very little nonresponse (less than 2% on almost all items) was not due to a structured skip pattern. In this analysis we treat all types of nonresponse identically.

Item response rates were lowest (as low as 4%) for problem items, several of which dealt with specialty services such as therapy or home health care needed by relatively few respondents. Some of the frequency and yes/no items also had low response rates. The greatest variation in the proportions of skipped items was evident among the yes/no items: 96.7% for a "complaint or problem

with plan" to 12.5% for "get prescription through plan". Domain mean outcomes are in general concentrated towards the higher end of their scales, indicating that most responses were favorable.

Table 1
Distribution of Responses and Ratings Evaluated over Items of the Same Type (n = 705,848 Respondents)

Statistic	Numerical	How Often	Problem	Yes/No
Number of items	4	11	11	9
Percentage responding				
Mean	74.97	62.56	30.32	57.26
Minimum	50.90	27.70	4.00	12.50
Maximum	95.00	74.50	64.40	96.70
Item means				
Mean	8.76	3.57	2.70	1.78
Minimum	8.57	3.09	2.49	1.62
Maximum	8.88	3.84	2.86	1.97
Distribution of ratings (across items in group)				
0	0.5			
1	0.4	2.0	5.7	19.5
2	0.4	6.3	12.1	80.5
3	0.7	23.9	82.2	
4	0.9	67.8		
5	4.6			
6	3.0			
7	6.2			
8	16.1			
9	17.8			
10	49.5			

Items are on a 0–10 numerical scale from "worse possible" to "best possible", a 4–point 1–4 ordinal "frequency" scale (never/sometimes/usually/always), a 3–point 1–3 ordinal "problem" scale (not a problem/somewhat a problem/a big problem), or are dichotomous 1–2 items (no/yes).

The domain mean, minimum, and maximum values across all items of the same type are also presented in Table 1. These illustrate that the 0–10 items have the smallest total variation (after rescaling to the common 0–1 range), while the 1–2 items have the largest total variation across domains and items. This is also illustrated in Figure 1, where we observe that the distribution of the 1–2 items varies substantially across items whereas the distributions of the 0–10 items are more homogeneous.

Table 2 presents statistical summary measures for the means and standard deviations of the domain mean ratings, evaluated across items of the same type. This complements Figure 1 by summarizing the difference in distributions of items within a given scale. Items with more response categories are concentrated towards the top of the scale and hence have smaller variance. For example, the mean standard deviation of the 1–2 items (0.36) is twice that of the rescaled 0–10 items (0.172). With the exception of the 0–10 items, the distributions of domain mean ratings vary greatly across items of the same type. For instance, the standard deviation of the means of 1–2 items across items is 0.30 compared to a rescaled standard deviation of 0.03 for the 0–10 items.

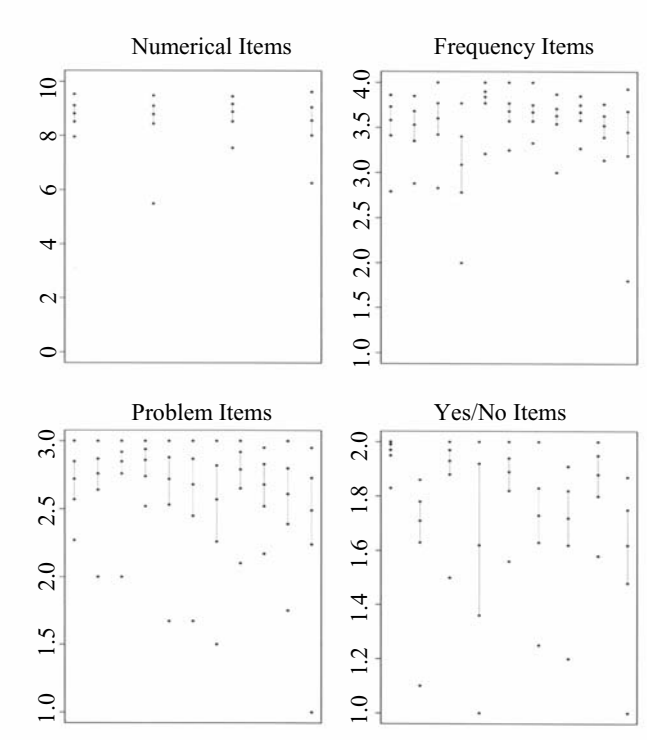


Figure 1. Five-point Summary of the Domain Sample Means for Each Item. The five-point summary consists of the minimum, 10th percentile, mean, 90th percentile, and the maximum.

Table 2
Summary Statistics of Domain Means and Standard Deviations Evaluated Over Domains and Items

Type	Summary Statistics for: Item Means				Item SDs	
	Min	Max	Mean	SD	Mean	SD
Numerical 0–10	6.82	9.52	8.76	0.30	1.72	0.26
Frequency 1–4	2.86	3.90	3.57	0.12	0.66	0.09
Problem 1–3	1.88	2.99	2.70	0.14	0.57	0.13
Yes/No 1–2	1.34	1.96	1.78	0.08	0.36	0.06

Note: Columns 2 through 5 give the minimum, maximum, mean, and standard deviation of the domain item means across items of a given type. Columns 6 and 7 give the mean and standard deviation of the domain item standard deviations across items of a given type.

Sample correlations also varied greatly across the pairs of items (Figure 2), although most were positive. Correlations between items of the same type most often were higher than those between items of different types. The numerical 0–10 ratings had the largest correlations (mean = 0.49), and generally ratings with more categories tended to have higher correlations than ratings with fewer categories. Although most of the pairs of 1–4 items had mean correlations near to 0.5, one item was negatively correlated with the others (revealed by the cluster of mean correlations below 0); this arose from reverse coding an item whose overall sample mean was not in the top half of the scale. The distributions

of the correlations of pairs of 1–2 items were centered near 0, indicating that pairs of items of this type often have negative correlations. Complete item wordings and additional summary statistics appear in Zaslavsky, Beaulieu, Landon and Cleary (2000) and Zaslavsky and Cleary (2002).

Models fitted to the variances and correlations are presented in the remainder of this section. Extensive checking of the best-fitting models indicated that the residuals did not follow any discernible pattern.

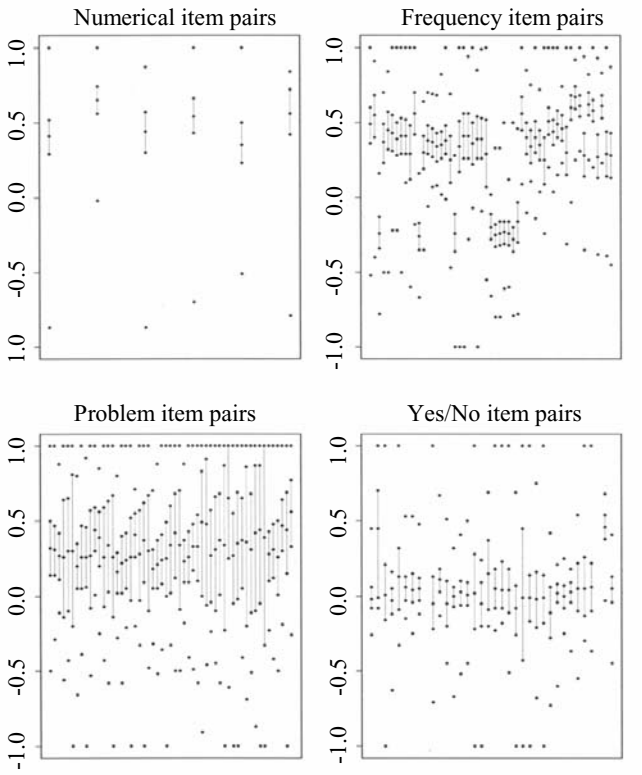


Figure 2. Five-point summary of the domain sample correlations between items with the same type. The five-point summary consists of the minimum, 10th percentile, mean, 90th percentile, and the maximum.

4.2 Variance Functions

In preliminary investigations not reported here, we fit two models within groups of items with the same response scale, one with common and one with different regression parameters for each item, to the data set comprising all of the items. Comparisons of the overall fits of the models (using criteria such as Mallow's C_p, R^2, adjusted R^2) and tests of the significance of effect-item interactions demonstrated that allowing parameters to vary across items significantly improved model fit. For instance, for the rescaled numerical ratings, weighted by domain sample size, the two models' root mean squared errors were 0.446 versus 0.402, and values of R^2 were 0.783 versus 0.825. Based on this we decided to fit separate models for each item.

The variance functions (8–10) were fitted to each item except the yes/no items, which follow the binomial variance function in the equal-probability sampling case. The iterative procedure described in section 3.4 converged almost precisely in exactly two iterations. This is because the weights for the observations change only with the estimate of τ^2, and so very little change in the weights occurs after the first iteration.

Table 3 presents the average sampling variation, average model error variation, and R^2, for each model averaged over items of each response scale. Sampling variation, computed using (19), does not depend on the model.

Table 3
Goodness-of-fit Statistics for Variance Functions

Rating Scale	0–10		1–4		1–3	
Sampling Variation	0.1460		0.3511		3.1703	
	ModErr	R^2	ModErr	R^2	ModErr	R^2
Model V1	0.020	0.741	0.066	0.824	0.069	0.916
Model V2	0.043	0.710	0.036	0.835	0.000	0.940
Model V3	0.016	0.750	0.024	0.847	0.000	0.947
	Prob(ModErr < Sampling Variation)					
Model V1	0.968		0.916		0.996	
Model V2	0.858		0.967		0.996	
Model V3	0.981		0.983		0.996	

ModErr is the variance component for lack of fit, R^2 is as defined in section 3.4, Prob(ModErr < Sampling Variation) is the proportion of domains for which model error is smaller than sampling variation. All ratings are rescaled to a 0–1 scale, and model errors are multiplied by 10^4.

For items with few categories (more closely resembling the binomial), the quadratic component of the variance function tends to dominate the linear component, making models V2 and V3 fit better than V1. Because V2 imposes a constraint at a point far outside the range of the domain means, it does not fit the data as well when there are more categories and the data are consequently further from binomial. The 0–10 items are less dispersed than the 1–4 and 1–3 ratings, enabling the linear model to fit better. The R^2 values for model V3 were close to 0.75 for numerical (0–10) items, 0.85 for the frequency (1–4) items, and 0.95 for the problem (1–3) items.

The lower portion of Table 3 displays for each item the proportion of domains (of those with at least 2 responses to the given item) for which sampling variation is larger than model error variation. For over 90% of domains, model error variation was less than the sampling variation of the direct variance estimate.

Figure 3 illustrates the fit of V3 for two each of the 0–10, 1–4, and 1–3 items. Illustrations for the remaining items are similar, but are not provided due to space limitations. The fitted curves are constrained to 0 at the maximum ratings. To assess the impact this constraint has on the fitted

variance function, we also fit an unrestricted (three parameter) quadratic variance function; these attained values very close to 0 at the maximum rating, and closely approximated the fitted curve from the constrained models, further supporting V3.

Average parameter estimates and their standard deviations over items of the same type are shown in Table 4. The parameters differed substantially across items, supporting the decision to estimate separate regression coefficients. In most cases the coefficients for both the $p_{h,i}$ and $p_{h,i}(1-p_{h,i})$ terms in V3 were significant, indicating that these are needed for generalized variance modeling. In some cases (particularly with the 0–10 items) the coefficient of the $p_{h,i}(1-p_{h,i})$ term was negative, resulting in an estimated variance function that is convex rather than concave (the shape of the binomial variance function). This can happen when the sample means for the ratings are concentrated on a small proportion of the response scale, over which the linear term explains much of the variation in the data. As mentioned earlier, adding higher-order polynomial or logarithmic functions of $p_{h,i}$ did not significantly improve model fit.

Table 4
Average Variance Function Parameter Estimates for Each Type of Item and Standard Deviations Across Items (in Parentheses)

Model	Item Type					
	0–10		1–4		1–3	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
V1	0.236	–	0.354	–	0.569	–
	(0.016)	–	(0.039)	–	(0.068)	–
V2	–	0.271	–	0.421	–	0.711
	–	(0.020)	–	(0.034)	–	(0.069)
V3	0.334	–0.114	0.151	0.241	0.239	0.420
	(0.143)	(0.155)	(0.104)	(0.132)	(0.112)	(0.110)

See Table 1 for a description of the 0–10, 1–4, and 1–3 items.

4.3 Correlation Functions

Models are ordered from simplest (C1, the constant model) to most complex (C5, containing all linear and quadratic terms). As for the variance models, statistical tests found highly significant item interaction effects, implying that separate models should be fit for each pair. We did not expect all pairs of items to have similar correlations, since by intention the items are divided into internally consistent groups, each of which measures a distinct aspect of patient experiences such as interactions with doctor or dealings with customer service agents (Hays, Shaul, Williams, Lubalin, Harris-Kojetin, Sweeny and Cleary 1999).

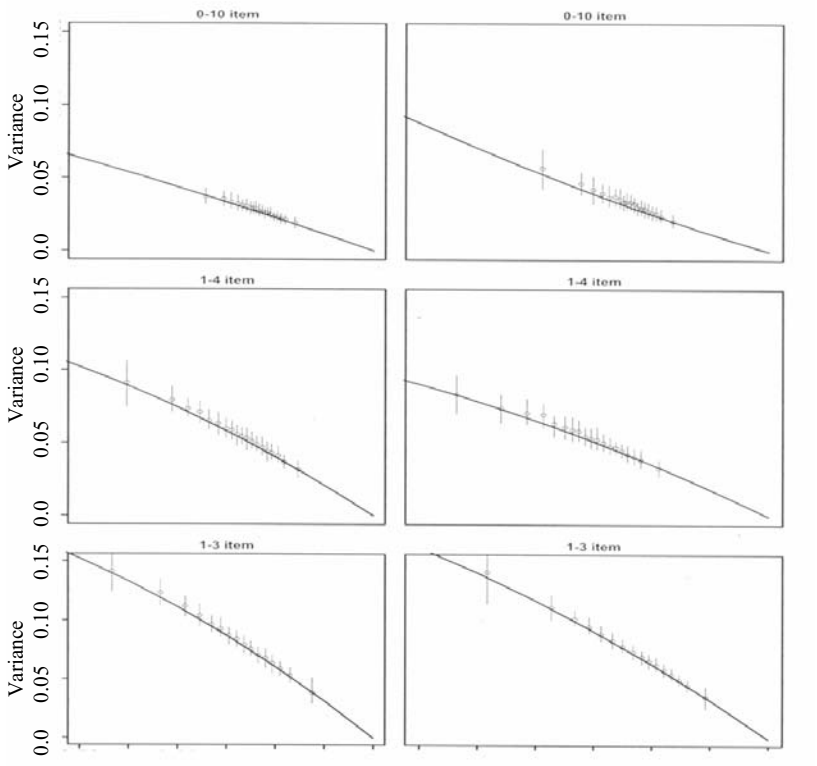


Figure 3. Quadratic Variance Function (V3) of Two Items for each Rating Type. Each point is the average of 60 domains. Vertical lines join the 10th and 90th percentiles of the distribution of the variances. For this and following displays the direction of the transformed horizontal axis has been reversed to agree with that of the original variables.

The fits of the correlation models for pairs of items of the same type are summarized in Table 5. Over the range of models considered, the biggest improvements in model performance (as measured by R^2) occur between model C1 and model C2, and between model C3 and model C4. For example, the average R^2 for the numerical ratings in models C3–C5 are 0.0391, 0.1494, and 0.1508 respectively, and the average R^2 for the 1–4 ratings over C1–C3 are 0, 0.0700, and 0.0789 respectively. This suggests that C2 and C4 are the best models for different pairs of items, a claim that is supported by the hypothesis tests on the significance of the incremental improvements in model fit.

Sampling variation was highest for the 1–3 ratings, at least in part because high rates of non-response due to skipped responses diminished the sample sizes. Model error and R^2 of correlation models for items of different types were similar to those for models for items having the same type.

The R^2 values for the correlation models were between 0.029 and 0.15 for all pairs of items. Although there was no evidence to suggest that C4 was an inappropriate model for the correlations, these results indicate that substantial variation in the correlations is not explained by the item means.

The sampling variances of the direct estimates were often less than the corresponding model error variances (lower part of Tables 5 and 6 especially for the 0–10 items. Under C4, model error variances were smaller for only 13% of domains for the 0–10 ratings, 45% of domains for the 1–4 ratings, and approximately 81% of domains for the 1–3 and 1–2 ratings.

Figure 4 presents the observed correlations and fitted function C4 for an illustrative pair of items from each of the 10 combinations of item types, representing the 595 distinct pairs of items. To illustrate the fitted correlation models, we adjust the observed and fitted correlations to the mean of one item and plot the resulting values in two-dimensional space. This process is repeated for the other item, yielding two plots for each correlation.

Figure 4 illustrates the generally weak relationship of the correlation to the means of the items seen in Tables 5 and 6. Analysis of Tables 5 and 6 reveals that the relationship between the correlation and the mean outcome is weaker for items with fewer categories and with correlations of items of different types. In particular, the 0 –10 numerical ratings are the only group for which there is a clear correlation-mean relationship.

Table 5
Model Fitting Diagnostics for Correlation Functions for Items of the Same Type, Averaged over Pairs of Items of the Same Type

Rating Type	0 – 10		1 – 4		1 – 3		1 – 2	
Sampling Variation	0.0124		0.0178		0.1482		0.0325	
	ModErr	R^2	ModErr	R^2	ModErr	R^2	ModErr	R^2
Model C1	0.060	0.000	0.028	0.000	0.112	0.000	0.018	0.000
Model C2	0.060	0.013	0.025	0.070	0.103	0.048	0.017	0.014
Model C3	0.057	0.039	0.024	0.079	0.102	0.054	0.017	0.018
Model C4	0.047	0.150	0.023	0.100	0.100	0.068	0.016	0.029
Model C5	0.044	0.151	0.023	0.105	0.096	0.080	0.015	0.034
	Prob(ModErr < Sampling Variation)							
Model C1	0.033		0.339		0.461		0.788	
Model C2	0.033		0.400		0.498		0.795	
Model C3	0.034		0.411		0.502		0.796	
Model C4	0.038		0.435		0.516		0.799	
Model C5	0.065		0.440		0.530		0.802	

See Table 1 for a description of the 0–10, 1–4, 1–3 and 1–2 items, and Table 3 for an explanation of the column headings.

Table 6
Model Fitting Diagnostics for Correlation Functions for C4 by Type of Item.
Averaged over Items of the Same Type

Types	0 – 10		1 – 4		1 – 3		1 – 2	
	ModErr	R^2	ModErr	R^2	ModErr	R^2	ModErr	R^2
0–10	0.047	0.149	0.021	0.104	0.040	0.094	0.013	0.059
1–4			0.023	0.100	0.038	0.076	0.013	0.039
1–3					0.100	0.068	0.028	0.031
1–2							0.016	0.029
Prob(ModErr < Sampling Variation)								
0–10	0.038		0.358		0.523		0.784	
1–4			0.435		0.605		0.790	
1–3					0.516		0.827	
1–2							0.799	

See Table 1 for a description of the 0–10, 1–4, 1–3 and 1–2 items, and Table 3 for an explanation of the column headings.

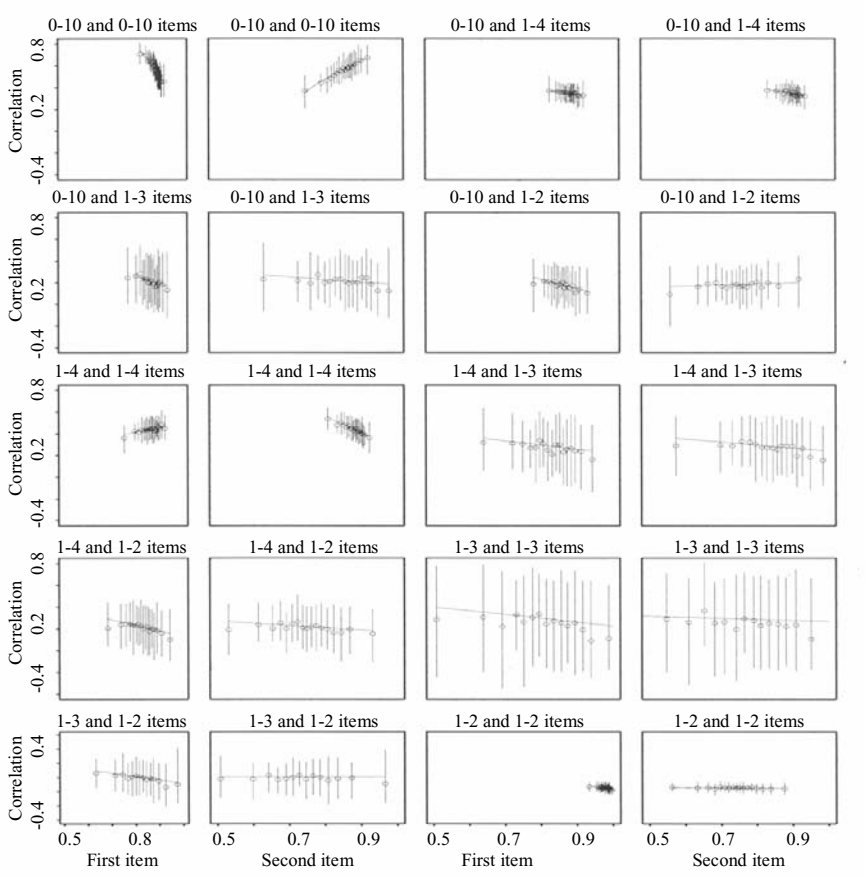


Figure 4. Correlation Functions for One Pair of Items for Each Combination of Rating Types.

Note: The plots for each items involved in the correlation are side by side. Refer to Figure 3 for a description of the contents and axes of the plot.

Although the fitted curves for the correlation functions are nearly flat, the variation in the parameter estimates under model C4 for α_4 are large and were suggestive of instability. The wildly varying parameter estimates are a consequence of collinearity among the predictors in model C4. In many cases the estimated value of α_4 offsets the parameter estimates for the linear predictors, resulting in a fitted curve that is nearly flat.

4.4 Mean Difference Functions

The difference $\hat{D}_{h,ij}$ appeared to depend on both the marginal mean and its square, implying a model analogous to V3 could be appropriate. However, because $\hat{D}_{h,ij}$ typically is small enough that $\hat{D}_{h,ij}\,\hat{D}_{h,ji}$ has minimal impact on (16), we fit a constant model.

4.5 Composite Estimator

Table 7 presents the quantiles of the distribution of weights $\sigma_h^2/(\tau^2+\sigma_h^2)$ for the model-based estimate, used in the composite estimator of section 3.5, averaged over items (or pairs of items) of the same type. The proportion of domains for which the standard error of the model-based predictions was smaller than that of the direct estimates is also presented. As noted previously, the model-based predictions have more weight in the composite variance estimates than in the composite correlation estimates. The average (across items or pairs) median of the weights of the model-based estimator ranged from 0.892 to 1.000 for variances, 0.256 to 0.709 for correlations of items of the same type, and from 0.468 to 0.738 for correlations of items of different types. Also, for both variances and correlations, the weight of the model-based predictions was larger for items with fewer response categories. For example, the model-based estimator had median weights of 0.256, 0.468, 0.540, and 0.647 on the composite estimates of correlations when the numerical 0–10 ratings were paired with the 0–10, 1–4, 1–3, and 1–2 ratings, respectively. However, even for pairs of 0–10 numerical ratings, for which sampling error of the direct estimator exceeded the model error in only 3.81% of domains, these results indicate that the median weight of the model-based estimator was 0.256, a nontrivial amount.

Table 7
Distribution of Weights for the Model-Based Component of the Composite Estimator, Averaged Over Items of Same Type

Model	Item Type		Prob(ModErr <	Quantiles		
	1	2	Sampling Variation)	10%	Median	90%
Variance	0 – 10	–	0.981	0.778	0.892	0.948
	1 – 4	–	0.983	0.948	0.966	0.974
	1 – 3	–	0.996	1.000	1.000	1.000
Correlation	0 – 10	0 – 10	0.038	0.141	0.256	0.335
	0 – 10	1 – 4	0.358	0.301	0.468	0.562
	0 – 10	1 – 3	0.523	0.357	0.540	0.654
	0 – 10	1 – 2	0.784	0.531	0.695	0.767
	1 – 4	1 – 4	0.435	0.324	0.497	0.591
	1 – 4	1 – 3	0.605	0.404	0.587	0.699
	1 – 4	1 – 2	0.853	0.584	0.738	0.805
	1 – 3	1 – 3	0.516	0.349	0.540	0.675
	1 – 3	1 – 2	0.827	0.584	0.737	0.817
	1 – 2	1 – 2	0.799	0.541	0.709	0.780

The distribution of weights is summarized by the 10^{th}, 50^{th}, and 90^{th} percentiles. See Table 3 for definition of ModErr.

4.6 Joint Predictions

Because we modeled the correlations independently for each item, our fitted correlation matrices do not necessarily satisfy the constraint of positive definiteness, which can be important for multivariate inference. In additional work, we have determined that as long as the multivariate analysis is restricted to items of the same type, the fitted correlations from the C2 and C4 models yield positive definite estimates of correlation matrices for almost all domains. However, for analyses including items of different types (*e.g.*, the 0–10 numerical items, and the 1–2 yes/no items), predictions based on C4 predict correlation matrices that are indefinite for many domains, while predictions based on C2 are more stable and almost always yield positive definite predictions. This suggests that while C4 may be slightly superior in terms of univariate model fit, C2 may be more appropriate for multivariate inference.

One way of overcoming the problem of indefinite predicted correlation matrices is to use a weighted average of the predicted correlation matrix for a domain and the estimated average correlation matrix (EACM) across domains. The EACM may be constructed by weighting the direct estimates (each of which is at least positive semi-definite) by the total sample size for each domain. Then any indefinite predicted correlation matrices are replaced with the weighted average of the predicted correlation matrix and the EACM, where the weight used for each domain is increased until a positive definite matrix results. Like an empirical Bayes estimator, this process stabilizes estimates by effectively shrinking the model coefficients toward those of a simpler (constant) model.

When analyzing all 35 CAHPS items simultaneously the EACM had an average weight across domains of 0.65 with model C4, whereas with model C2 the average weight was only 0.01 since the predicted correlations under C2 were usually positive definite. In analyzing only the 0–10, 1–4, and 1–3 items the EACM had average weights of 0.28 and 0.00 with C4 and C2 respectively, while in analyzing just the 0–10 and 1–4 items the corresponding average weights were 0.06 and 0.00. When analyzing the different types of items separately, the average weight of the EACM with C4 was 0.00 for the 0–10 and 1–4 items, 0.01 for the 1–3 items, and 0.17 for the 1–2 items. The EACM is thus not needed when analyzing the 0–10 and 1–4 items because the predicted correlation matrices were positive definite for every domain.

5. Conclusion

We have presented methodology for estimating variance and covariance functions for domain means of ordinal survey items. Our methodology can also be applied to survey items measured on continuous scales. We introduced a decomposition of the model error that allows the variation due to sampling to be separated from that due to model fit. The decomposition also helps to avoid over-fitting because it estimates the proportion of variation in the data that can be modeled and thus when the current predictors suffice.

The procedure for fitting the variance and correlation models is the same regardless of whether or not the data contain skip patterns. The analytic derivation in section 3.3 shows that if skip patterns are present, mean differences of items by response status of other items are required in order to compute the sampling covariance estimates. However, we argued that these quantities are likely to have minimal impact on the results and that therefore a constant model

could be used, which was supported by our empirical findings.

A quadratic variance function constrained to 0 at the maximum rating, and a model for transformed correlations involving the product but not the squares of the means, best predicted the direct estimates in our applied example. The modeled variance estimates generally had much smaller standard errors than the direct estimates; the same was, however, not true of the correlation estimates. It is interesting and reassuring that our quadratic variance function can be expressed as the widely-used relative variance model of Wolter (1985).

For our ordinal data, the estimates of the domain mean ratings contain minimal information about the correlation between the ratings. Hence, the mean-covariance relationship is principally an artifact of the mean-variance relationship. However, for items with many response categories, the association between correlations and mean outcomes for items of the same type was stronger most notably for pairs of 0–10 items. With the exception of the 0–10 and possibly the 1–4 ratings, the correlations might as well be modeled as constants, which also makes it easier to guarantee positive definiteness of the predicted correlation matrix. However, it is important that the parameters of the correlation model be allowed to vary across pairs of items.

A composite estimator that weights the direct and model-based estimators proportional to their precisions has smaller variance than either estimator alone, especially when the components have close to equal weight. The model-based estimator had the greatest influence on estimates for small domains, for which little information is available. The model-based estimator had the greatest influence on estimates for variances, followed by correlations of items of the same type, and lastly correlations of items of different types. Both model-based and composite estimators can be benchmarked (ratio adjusted) to agree on the average across domains with direct estimates, although this proved to be unnecessary in our example.

GVCFs find several applications in our continuing research. We are developing quasi likelihood-based methods for estimating covariance matrices for the domain means of ordinal survey items, representing the second-level (structural) covariance in a hierarchical model (O'Malley and Zaslavsky 2004). GVCF models are needed to provide estimates of sampling variances and covariances and to modify those estimates as the means are re-estimated during the fitting procedure. If the sampling variability of the GVCF estimates is minimal because the number of domains is large, the GVCF predicted variances and covariances can be treated as known. However, if the sampling error of the GVCF-based estimates is large a model that allows these errors to propagate through the analysis should be used. In related work, Fay and Train (1997) used a binomial model with a design effect for each domain in empirical Bayes estimation of binomial rates. Our research extends this approach to multivariate estimation and more general response formats.

Another application of GVCFs is the computation of variance estimates for linear combinations of item means, facilitating variance estimation for composite scores, like those used in CAHPS reporting. The methods described in section 2 are applicable to variance estimation for any functions of totals, including functions of means, other ratios, or regression coefficients.

There are several ways of extending the GVCF methodology. In addition to summary measures of outcomes, generalized variance and covariance functions (GVCFs) may also depend on other independent variables, in particular those that would better predict correlations. We considered variables summarizing response patterns, such as the proportion of respondents in a domain, but these did not improve the model. GVCFs could also be extended to multi-stage sampling.

Acknowledgements

This work was supported by the U.S. Agency for Healthcare Research and Quality through the Consumer Assessments of Health Plans Study (grant U18 HS09205-06) and by the U.S. Centers for Medicare and Medicaid Services (contract 500-95-007). We thank Paul D. Cleary for his ongoing support of this work, Matt Cioffi for data management, and Elizabeth Goldstein and Amy Heller of the Centers for Medicare and Medicaid Services (CMS), and the other members of the CAHPS-MMC survey implementation team.

References

Cho, M.J., Eltinge, J.L., Gershunskaya, J. and Huff, L.L. (2002). Evaluation of generalized variance function estimators for the U.S. Current Employment Survey. In *Proceedings of the Joint Statistical Meetings* [CDROM]. Alexandra, VA: American Statistical Association, 534-539.

Eltinge, J. (2002). Use of generalized variance functions in multivariate analysis. In *Proceedings of the Section on Survey Research Methods*, Alexandra, VA: American Statistical Association, 904-913.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Fay, R.E., and Train, G.F. (1997). Small domain methodology for estimating income and poverty characteristics for states in 1993. In *Proceedings of the Social Statistics Section*, Alexandra, VA: American Statistical Association, 183-188.

Freund, J.E., and Walpole, R.E. (1987). *Mathematical Statistics*. New Jersey: Prentice-Hall, Inc., 4th Edn.

Gabler, S., Haeder, S. and Lahiri, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, 25, 105-106.

Goldstein, E., Cleary, P.D., Langwell, K.M. Zaslavsky, A.M. and Heller, A. (2001). Medicare Managed Care CAHPS: A tool for performance improvement. *Health Care Financing Review*, 22, 101-107.

Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-76.

Hays, R.D., Shaul, J.A., Williams, V.S.L., Lubalin, J.S., Harris-Kojetin, L.D., Sweeny, S.F. and Cleary, P.D. (1999). Psychometric properies of the CAHPS 1.0 survey measures. *Medical Care*, 37 (Supplement), 22-31.

Huff, L.L., Eltinge, J.L. and Gershunskaya, J. (2002). Exploratory analysis of generalized variance function models for the U.S. Current Employment Survey. In *Proceedings of the Joint Statistical Meetings* [CDROM], Alexandra, VA: American Statistical Association, 1519-1524.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons, Inc.

O'Malley, A.J. and Zaslavsky, A.M. (2004). Implementation of cluster-level covariance analysis for survey data with structured nonresponse. In *Proceedings of the Joint Statistical Meetings*, American Statistical Association, 1907-1914.

Otto, M.C., and Bell, W.R. (1995). Sampling error modeling of poverty and income statistics for states. In *Proceedings of the Section on Government Statistics*, American Statistical Association, 160-165.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Seber, G.A.F. (1977). *Linear Regression Analysis*. New York: John Wiley & Sons, Inc.

Spencer, B.D. (2000). An approximate design effect for unequal weighting when measurements may correlate with selection probabilities. *Survey Methodology*, 26, 137-138.

Valliant, R. (1992a). Longitudinal smoothing of price index variances. In *Statistics Canada Symposium*. Ottawa: Statistics Canada. 113-120.

Valliant, R. (1992b). Smoothing variance estimates for price indexes over time. *Journal of Official Statistics*, 8, 433-444.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference*. New York: John Wiley & Sons, Inc.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Woodruff, S. (1992). Variance estimation for estimates of employment change in the Current Employment Statistics Survey. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA: American Statistical Association, 626-631.

Zaslavsky, A.M., Beaulieu, N.D., Landon, B.E. and Cleary, P.D. (2000). Dimensions of consumer-assessed quality of Medicare managed-care health plans. *Medical Care*, 38, 162-174.

Zaslavsky, A.M., and Cleary, P.D. (2002). Dimensions of plan performance for sick and healthy members on the Consumer Assessments of Health Plans Study 2.0 survey. *Medical Care*, 40, 951-964.

.