THE ANALYSIS OF STRATIFIED MULTIPLE RESPONSES

Ivy Liu and Thomas Suesse
School of Mathematics, Statistics and Computer Science
Victoria University of Wellington
New Zealand

October 17, 2007

SUMMARY

Surveys often contain qualitative variables for which respondents may select any number of the outcome categories. For instance, for the question "What type of contraception have you used?" with possible responses (oral, condom, lubricated condom, spermicide, and diaphragm), respondents would be instructed to select as many of the outcomes that apply. This situation is known as *multiple responses*. When the data includes stratification variables, we discuss two approaches: (1) the model-based approach which uses logit models directly applying the generalized estimating equations (GEE) method (Liang and Zeger, 1986); and (2) the non-model-based approach which extends the generalized Mantel-Haenszel type estimators (Greenland, 1989) to make inferences across multiple responses. These approaches can also be used for data with dependent observations across strata.

Key words: bootstrap method, dually consistent, generalized estimating equations, generalized Mantel-Haenszel estimator, multiple responses; odds ratio.

1 Introduction

In many surveys, it is common that respondents may select any number of the outcome categories. For instance, for the question "What type of contraceptives have you used?" with possible responses (oral, condom, lubricated condom, spermicide, and diaphragm), respondents would be instructed to select whichever of the outcomes apply. Categorical variables that summarize this kind of data are called *pick any/m variables* (Bilder and Loughin, 2002), where m is the number of outcome categories ($m = 5$ in this case). We can cross-classify the counts from a survey that contains a pick any/m variable along with a group variable (r levels, e.g. whether a subject had a prior history of urinary tract infection) and a stratification variable (K levels, e.g. several age groups) into an $r \times m \times K$ contingency table. In the $r \times m \times K$ table, subjects may be represented in more than one cell. Data for such an example for 239 sexually active college women in a $2 \times 5 \times 2$ table is given in Table 1 by Bilder and Loughin (2002). We are interested in the conditional relationship between the type of contraception and a prior history of urinary tract infection given the age groups.

Table 1: The marginal UTI data

	Contraceptive					Total	Total
	Oral	Condom	L. cond.	Spermicide	Diaphragm	responses	women
Age $\geq$ 24							
UTI							
No	18	9	8	7	0	42	24
Yes	8	9	2	3	2	24	14
Age $<$ 24							
UTI							
No	55	41	37	27	0	160	85
Yes	75	68	33	22	5	203	116

Another example comes from a study conducted by Dr. Paul Warren in the School of Linguistics and Applied Language Studies at Victoria University of Wellington, New Zealand. The data was generated by 6 experts (raters) rated 50 non-native English utterances into 3 scales for overall comprehensibility (from "not easy" to "very easy" to understand) and then indicated whether there was a problem for that utterance in each of 7 items (e.g. pronunciation of consonants, vowel pronunciation, word stress, etc.). These 7 items are the pick any/m variables, where $m = 7$ in this example. Each item can be treated as a binary choice (i.e., it was or was not a problem). The study was interested in evaluating the conditional relationship between the overall rating and the 7 items given the raters. Table 2 shows 6 separate 3×7 tables ($K = 6$, $r = 3$, and $m = 7$), where the cell counts are dependent across the columns for each table and also dependent across the 6 strata.

Both examples are of stratified multiple response data, yet the observations are not independent across the strata in the second example. This type of data occurs frequently in health and social sciences, and in language studies. To analyze the data, we need the complete information on which items have been selected for each of the women (Example 1) or utterances (Example 2). One can express the complete information for each of the respondents using an $r \times 2^m \times K$ contingency table as in Table 3, where the columns form the response profile on the m items. In total, there are 2^m possible profiles, according to the (yes, no) outcome for the selection of each item. The complete information on each of the 50 utterances on the 6

Table 2: The marginal Linguistics data

	Items							Total	Total
	1	2	3	4	5	6	7	responses	utterances
Rater 1									
Rating									
1	8	7	2	2	1	0	1	21	8
2	32	22	7	2	6	0	3	72	32
3	8	1	3	0	0	0	1	13	10
Rater 2									
Rating									
1	10	8	8	4	5	8	0	43	11
2	18	6	10	11	8	11	1	65	19
3	18	9	4	3	8	7	0	49	20
Rater 3									
Rating									
1	7	1	3	0	4	2	0	17	7
2	11	4	6	1	8	4	0	34	13
3	23	7	8	3	13	8	2	64	30
Rater 4									
Rating									
1	2	2	2	2	0	0	0	8	2
2	11	7	2	4	1	1	0	26	12
3	11	6	1	5	0	0	1	24	36
Rater 5									
Rating									
1	1	0	0	0	0	0	0	1	1
2	8	6	5	0	1	1	0	21	23
3	5	11	4	0	1	1	0	22	26
Rater 6									
Rating									
1	14	18	6	14	14	17	0	83	18
2	12	10	1	9	11	9	0	52	14
3	12	14	4	7	9	11	1	58	18

raters and 7 items can be displayed in a similar fashion.

Table 3: The complete UTI data

Age $\geq$ 24																
Contraceptive																
Oral	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Condom	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
L. cond.	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
Spermicide	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
Diaphragm	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
UTI																
No	0	0	0	0	0	0	0	0	0	0	0	0	2	0	4	0
Yes	0	0	0	0	0	0	0	0	2	0	1	1	1	0	0	1
Contraceptive																
Oral	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Condom	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
L. cond.	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
Spermicide	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
Diaphragm	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
UTI																
No	14	0	1	0	0	0	0	0	1	0	0	0	0	0	2	0
Yes	5	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0
Age $<$ 24																
Contraceptive																
Oral	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Condom	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
L. cond.	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
Spermicide	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
Diaphragm	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
UTI																
No	0	0	1	0	0	0	0	0	2	0	1	0	8	0	18	0
Yes	0	0	1	0	0	0	0	0	14	0	3	0	10	0	12	1
Contraceptive																
Oral	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Condom	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
L. cond.	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
Spermicide	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
Diaphragm	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
UTI																
No	42	0	1	0	0	0	0	0	1	0	0	0	5	0	6	0
Yes	44	3	0	0	0	0	0	0	15	1	2	0	7	0	3	0

Bilder and Loughin (2002) generalized the Cochran test to determine if the group and pick any/m variable (or "items") are marginally independent given a stratification variable (known as conditional multiple marginal independence, CMMI). For the UTI example, they tested whether the contraception practices of women are different based on their urinary tract infection history controlling for their age group. They used a nonparametric bootstrap method to obtain the p-value of the test. When the group and items are not conditionally marginally independent, it is more interesting to describe how the items depend on the group. Similarly, for the Linguistics example, we are not interested in the differences between raters, and we focus on describing the conditional relationship between the overall rating and the items given each rater.

This article discusses two approaches to the analysis of such data. The first approach, called the *model-based* approach, treats the m items as a m-dimensional binary response

and then uses logit models directly for the marginal distribution of each item. It applies the methodology of generalized estimation equations (GEE) (Liang and Zeger, 1986) that is a multivariate extension of quasi-likelihood methods. With the GEE method, we need to provide structure only for how the variance depends on the mean and the correlation structure of the m items. Besides the GEE method, Agresti and Liu (1999, 2001) have provided other methods to fit the models, such as the Maximum Likelihood (ML) method. The GEE method is the computationally simplest one. The second approach, called *non-model-based* approach, extends the generalized Mantel-Haenszel (MH) estimators (Greenland, 1989) to make the inference across m items. The MH-type estimators have been used in many cases involving stratified data (Liu and Agresti, 1996; Liu, 2003). The MH-type estimators are *dually consistent*, that is, the estimators are consistent under two types of asymptotics – (a) when the sample size within each stratum increases and the number of strata is fixed, and (b) when the number of strata increases proportional to the overall sample size. Sparse data fall into the type (b) situation. For an ordinary binary response case, it is well known that the MH estimators perform much better than the ML estimators for sparse data (Andersen, 1980, p. 244). To make the inference across m items, we derive the dually consistent variance and covariance estimators for the generalized MH estimators. Similar to the Cochran-Mantel-Haenszel test, generalized MH estimators are used when the conditional associations are not expected to vary drastically among the strata. However, even though the true associations are heterogeneous between strata, the generalized MH estimators often provide a useful descriptive summary if the directions of the associations are the same across strata.

Section 2 introduces the model-based approach using the GEE method. Section 3 shows the way that the generalized MH estimators apply to multiple responses and gives dually consistent variance and covariance estimators. Section 4 provides the data analysis for the two examples. The dually consistent variance and covariance estimators for the generalized MH estimators are applicable only when the strata are independent. When the strata are dependent as in the Linguistics example, it is more realistic to use the bootstrap method to evaluate the variance and covariances of the estimators, because the dually consistent ones are too complicated to derive. Therefore, in Section 5, we discuss the simulation results for the performance of the bootstrap method when the data are simulated from various situations. We also compare the performance between the GEE and MH methods. The last section provides a general discussion.

2 Model Based Approach

Consider the m items as a m-dimensional binary response. For each item, the response is either "the item is selected" or "the item is not selected". For example, for Linguistics data, we let $\pi_{j|ik}$ be the probability of having a problem on item j when the utterance is overall rated on level i by rater k. To describe our main interest about the conditional relationship between the overall rating and the items given each rater, we use the logit model for the marginal probabilities of each item having the form

$$\log\left(\frac{\pi_{j|ik}}{1-\pi_{j|ik}}\right) = \beta_{ij} + \tau_{jk}\,, \tag{1}$$

where $i = 1, \ldots, r$, $j = 1, \ldots, m$, and $k = 1, \ldots, K$. Identifiability requires constraints such as $\beta_{rj} = 0$ and $\tau_{jK} = 0$ for all j. Define $\gamma^j_{ab} = \beta_{aj} - \beta_{bj}$. The parameters $\{\gamma^j_{ab}\}$ characterize the conditional relationships. For instance, the odds of having a problem on item j when the utterance is overall rated on level a are $\exp(\gamma^j_{ab})$ times the odds of having a problem on item j when the utterance is overall rated on level b, given each rater. Since the GEE method is a multivariate extension of quasi-likelihood methods, we do not need to specify the full joint

distribution of m items. It only needs the structure for how the variance depends on the mean and the correlation structure of the m items. One can make a " working guess" about the correlation structure of the item responses and then adjust the standard error of the parameter estimators to reflect what actually occurs for the sample data using a "sandwich" method. In SAS, PROC GENMOD (the procedure for generalized linear models) with REPEATED statement can implement the GEE method for a variety of working correlation structures for the dependence among the m items.

The GEE approach is easy to apply for the UTI data, because the responses are dependent only across the m items and the observations are independent across strata. For the Linguistics data, it is not clear how the correlation structure can be chosen when the responses are correlated across the m items and also the K strata (raters), although one could always choose an "independent" working correlation structure and use the sandwich standard errors to take into account the empirical situation.

Instead of using the logit model, the conditional associations can also be obtained using a generalized MH-type estimator. Unlike the logit model, the MH-type method cannot be used to select the best model that includes all significant predictors. However, if one is particularly interested in the conditional association between the item and the overall rating given each rater, the MH-type estimators evaluate the association directly. The next section gives the details.

3 Non-Model Based Approach

Consider each item separately. For example, we consider item "1" (consonant pronunciation) only in Table 2. The conditional association between overall rating and "whether there was a consonant pronunciation problem" given the rater can be described using a $3 \times 2 \times 6$ table, where the column variable is "whether there was a consonant pronunciation problem" with two levels (yes, no), the row variable is overall rating (not easy, medium, very easy), and the stratum variable is rater. Suppose we naively treat the 3×2 tables for 6 raters as independent. We can use the generalized MH estimators (Greenland, 1989) to describe the conditional relationship between the row and column variables. The estimators themselves are consistent. However, the standard error and covariance estimates for the estimators based on the naive independent assumption are inappropriate. There are two ways to find proper standard errors and covariance estimates: (1) deriving dually consistent estimators; and (2) using the bootstrap method. We will discuss these in sections 3.1 and 3.2.

For a general $r \times m \times K$ table, let $X_{j|ik}$ denote the number of utterances having a problem on item j rated by the kth rater (stratum) with the overall rating (row) i. The notation n_{ik} denotes the total number of utterances in the ith row and the kth stratum. Let $N_k = n_{1k} + \ldots + n_{rk}$. For convenience, we also let $\bar{\pi}_{j|ik} = 1 - \pi_{j|ik}$ and let $\bar{X}_{j|ik} = n_{ik} - X_{j|ik}$. Define a common odds ratio for rows a and b as

$$\Psi^j_{ab} = \frac{\pi_{j|ak}\bar{\pi}_{j|bk}}{\bar{\pi}_{j|ak}\pi_{j|bk}} \quad j = 1, \ldots, m, \;\; a = 1, \ldots, r, \;\; b = 1, \ldots, r, \text{ and } a \neq b,$$

for all k. The Ψ^j_{ab} is the ratio of the odds of having a problem on item j for utterances overall rated a to the odds of having a problem on item j for utterances overall rated b, given any stratum. The generalized MH estimator (Greenland, 1989) of $\log \Psi^j_{ab}$ is

$$\bar{L}^j_{ab} = (L^j_{a+} - L^j_{b+})/r\,,$$

where $L^j_{ab} = \log\left(\frac{\sum_{k=1}^K X_{j|ak}\bar{X}_{j|bk}/N_k}{\sum_{k=1}^K X_{j|bk}\bar{X}_{j|ak}/N_k}\right)$ and the subscript "+" indicates summation over that subscript. When the row variable has only two levels ($r = 2$) as for the UTI example, we

can use Ψ_{12}^j to describe the conditional row effect on selecting item j. The generalized MH estimator of $\log \Psi_{12}^j$ is simplified to the ordinary MH estimator

$$L_{12}^j = \log\left(\frac{\sum_{k=1}^K X_{j|1k}\bar{X}_{j|2k}/N_k}{\sum_{k=1}^K X_{j|2k}\bar{X}_{j|1k}/N_k}\right).$$

3.1 Dually consistent variance and covariance estimators

When the strata are independent (e.g., UTI example), we can derive the dually consistent variance and covariance estimators for the generalized MH estimators. Suppose one is only interested in a particular item, say x ($x \in \{1,\ldots,m\}$), the dually consistent variance and covariance estimators for $\{\bar{L}_{ab}^x, \forall\, a \neq b\}$ are applicable directly from the work by Greenland (1989). However, one might be interested in comparing the conditional association across items. For instance, for the UTI example, one might be interested in comparing the UTI effects for contraceptive methods "oral" with "condom". The covariance estimator between $\bar{L}_{ab}^x$ and $\bar{L}_{ab}^y$ is desirable for $x \neq y$ ($x, y \in \{1,\ldots,m\}$). The way to derive the dually consistent estimator for it is more complicated than the case considering only a fixed item, because $X_{x|ik}$ and $X_{y|ik}$ are correlated for all i and k. That is, the numbers of women who used contraceptive methods x and y are not independent. To find the dually consistent covariance estimator, we need to consider up to the fourth moment of X's and the pairwise counts for the two items.

Define pairwise probabilities for items x and y ($x, y \in \{1,\ldots,m\}$) as $\pi_{xy|ik}^{b_1b_2}$ with $b_1, b_2 \in \{0,1\}$, where $(0,1)$ is the (no, yes) outcome for the selection of each item. The $\pi_{xy|ik}^{b_1b_2}$ is the probability of observing the pairwise outcome (b_1, b_2) for items x and y. For instance, the notation $\pi_{xy|ik}^{11}$ is the probability that a subject, who is in row i and stratum k, selects both items x and y. We assume that the pairwise probabilities $\pi_{xy|ik}^{00}$, $\pi_{xy|ik}^{01}$, $\pi_{xy|ik}^{10}$, $\pi_{xy|ik}^{11}$ follow a multinomial distribution. We have $\pi_{xy|ik}^{00}+\pi_{xy|ik}^{01}+\pi_{xy|ik}^{10}+\pi_{xy|ik}^{11} = 1$ and $\pi_{x|ik} = \pi_{xy|ik}^{10}+\pi_{xy|ik}^{11}$, $\pi_{y|ik} = \pi_{xy|ik}^{01} + \pi_{xy|ik}^{11}$. Define similarly the pairwise observations as $\{X_{xy|ik}^{b_1b_2}\}$.

First, we consider the fixed item case, say item x. Define $C_{x|ab} = \sum_{k=1}^K c_{x|abk}$ with $c_{x|abk} = X_{x|ak}\bar{X}_{x|bk}/N_k$, $h_{x|ab} = (X_{x|ak}+\bar{X}_{x|bk})/N_k$. Greenland (1989) derived the following estimators:

$$U_{x|ab} := \widehat{\mathrm{Var}}(L_{ab}^x) \quad U_{x|abc} := \widehat{\mathrm{Cov}}(L_{ab}^x, L_{ac}^x)$$

with

$$\begin{aligned} U_{x|ab} &= \frac{\sum_k c_{ab}h_{ab}}{2C_{ab}^2} + \frac{\sum_k c_{ba}h_{ab}+c_{ab}h_{ba}}{2C_{ab}C_{ba}} + \frac{\sum_k c_{ba}h_{ba}}{2C_{ba}^2} \\ U_{x|abc} &= \frac{\sum_k X_a\bar{X}_b\bar{X}_c/N_k^2}{3C_{ab}C_{ac}} + \frac{\sum_k n_a\bar{X}_bX_c/N_k^2}{3C_{ab}C_{ca}} + \frac{\sum_k n_aX_b\bar{X}_c/N_k^2}{3C_{ba}C_{ac}} + \frac{\sum_k \bar{X}_aX_bX_c/N_k^2}{3C_{ba}C_{ca}} \end{aligned}$$

For convenience, we often suppress subscripts x and k. For instance, $c_{ab} = c_{x|abk}$, $X_a = X_{x|ak}$, and $n_a = n_{ak}$.

Because $\bar{L}_{ab}^x$ is a linear combination of $\{L_{ab}^x\}$, $\widehat{\mathrm{Cov}}(\bar{L}_{ab}^x, \bar{L}_{cd}^x)$ can be expressed as follows in terms of $U_{x|ab}$ and $U_{x|abc}$:

$$\widehat{\mathrm{Cov}}(\bar{L}_{ab}^x, \bar{L}_{cd}^x) = (U_{x|ac}^+ - U_{x|ad}^+ - U_{x|bc}^+ + U_{x|bd}^+)/r^2 \tag{2}$$

with

$$U^{+}_{x|ab} = \begin{cases} U_{x|a++} = \sum_{i,k} U_{x|aik} & \text{for } a = b \\ U_{x|+ab} - U_{x|ab+} - U_{x|ba+} + U_{x|ab} = U^{+}_{x|ba} & \text{for } a \neq b \end{cases}$$

The subscript "+" denotes summation over that subscript. Note that setting $c = a$, $d = b$ yields $\widehat{\mathrm{Var}}(\bar{L}^x_{ab})$ and setting $c = a$, $d = c$ yields $\widehat{\mathrm{Cov}}(\bar{L}^x_{ab}, \bar{L}^x_{ac})$.

Next, we make the inference across two different items. For instance, consider the dually covariance estimator between $\bar{L}^x_{ab}$ and $\bar{L}^y_{cd}$. We propose the following dually consistent estimators:

$$\begin{aligned} U_{xy|ab} &:= \widehat{\mathrm{Cov}}(L^x_{ab}, L^y_{ab}) = \frac{\hat{D}_{ab}}{C_{x|ab}C_{y|ab}} - \frac{\hat{D}_{x|ab}}{C_{x|ba}C_{y|ab}} - \frac{\hat{D}^y_{ab}}{C_{x|ba}C_{y|ba}} + \frac{\hat{D}^{xy}_{ab}}{C_{x|ba}C_{y|ba}} \\ U_{xy|abc} &:= \widehat{\mathrm{Cov}}(L^x_{ab}, L^y_{ac}) = \frac{\hat{D}_{abc}}{C_{x|ab}C_{y|ac}} - \frac{\hat{D}^x_{abc}}{C_{x|ba}C_{y|ac}} - \frac{\hat{D}^y_{abc}}{C_{x|ab}C_{y|ca}} + \frac{\hat{D}^{xy}_{abc}}{C_{x|ba}C_{y|ca}} \end{aligned}$$

with $\hat{D} = \sum_k \hat{d}_k$ and

$$\begin{aligned} \hat{d}_{ab} &= \frac{1}{N_k^2}\{X_{x|a}X_{y|a}X^{00}_{xy|b} + X^{11}_{xy|a}\bar{X}_{x|b}\bar{X}_{y|b} - X^{11}_{xy|a}X^{00}_{xy|b}\} \\ \hat{d}^x_{ab} &= \frac{1}{N_k^2}\{\bar{X}_{x|a}X_{y|a}X^{10}_{xy|b} + X^{01}_{xy|a}X_{x|b}\bar{X}_{y|b} - X^{01}_{xy|a}\pi^{10}_{xy|b}\} \\ \hat{d}^y_{ab} &= \frac{1}{N_k^2}\{X_{x|a}\bar{X}_{y|a}X^{01}_{xy|b} + X^{10}_{xy|a}\bar{X}_{x|b}X_{y|b} - X^{10}_{xy|a}X^{01}_{xy|b}\} \\ \hat{d}^{xy}_{ab} &= \frac{1}{N_k^2}\{\bar{X}_{x|a}\bar{X}_{y|a}X^{11}_{xy|b} + X^{00}_{xy|a}X_{x|b}X_{y|b} - X^{00}_{xy|a}X^{11}_{xy|b}\} \end{aligned}$$

and

$$\begin{aligned} \hat{d}_{abc} &= \frac{1}{N_k^2}X^{11}_{xy|a}\bar{X}_{x|b}\bar{X}_{y|c} \\ \hat{d}^x_{abc} &= \frac{1}{N_k^2}X^{01}_{xy|a}X_{x|b}\bar{X}_{y|c} \\ \hat{d}^y_{abc} &= \frac{1}{N_k^2}X^{10}_{xy|a}\bar{X}_{x|b}X_{y|c} \\ \hat{d}^{xy}_{abc} &= \frac{1}{N_k^2}X^{00}_{xy|a}X_{x|b}X_{y|c} \end{aligned}$$

The formula, $U_{xy|ab}$ is invariant under interchange of items (x and y) or rows (a and b). Note that $U_{x|ab} = U_{x|ba}$, because $L^x_{ab} = -L^x_{ba}$. Also, $U_{x|abc} = U_{x|acb}$ due to the definition. However, $U_{xy|ab} \neq U_{xy|ba}$, but $U_{xy|ab} = U_{yx|ba}$ by the definition.

Again, Since $\bar{L}^x_{ab}$ (or $\bar{L}^y_{ab}$) is a linear combination of $\{L^x_{ab}\}$ (or $\{L^y_{ab}\}$), we can derive covariance estimators for $(\bar{L}^x_{ab}, \bar{L}^y_{cd})$, which can be expressed as follows:

$$\widehat{\mathrm{Cov}}(\bar{L}^x_{ab}, \bar{L}^y_{cd}) = \frac{1}{r^2}\{U^{+}_{xy|ac} - U^{+}_{xy|ad} - U^{+}_{xy|bc} + U^{+}_{xy|bd}\} \tag{3}$$

with

$$U^{+}_{xy|ac} = \begin{cases} U_{xy|a++} = \sum_{i,k} \mathrm{Cov}(L^x_{ai}, L^y_{ak}) & \text{for } a = c \\ U_{xy|+ac} - U_{xy|a+c} - U_{xy|ca+} + U_{xy|ac} & \text{for } a \neq c \end{cases}$$

For non-distinct indices a, b, c, d we have

$$\widehat{\mathrm{Cov}}(\bar{L}^x_{ab}, \bar{L}^y_{ac}) = \frac{1}{r^2}\{U^{+}_{xy|a++} - U^{+}_{xy|ac} - U^{+}_{xy|ba} + U^{+}_{xy|bc}\}$$

and

$$\widehat{\mathrm{Cov}}(\bar{L}^x_{ab}, \bar{L}^y_{ab}) = \frac{1}{c^2}\{U_{xy|a++} - U^{+}_{xy|ab} - U^{+}_{xy|ba} + U_{xy|b++}\}.$$

The Appendix provides dually consistency arguments for these estimators. We will refer later to "formulae" variance and covariance estimators, meaning Greenland's dually consistent variance $\widehat{\mathrm{Var}}(\bar{L}^x_{ab})$ in (2) and the dually consistent covariance $\widehat{\mathrm{Cov}}(\bar{L}^x_{ab}, \bar{L}^y_{cd})$ in (3) proposed above.

When the strata are not independent (e.g., Linguistic example), it is even more complicated to derive the dually consistent variance and covariance estimators, because the X's are correlated across not only items, but also strata. For instance, for $K = 6$, we need to consider up to the 24^{th} moment of X's. Because of this complexity, Section 3.2 provides a realistic way to find estimates by applying the nonparametric bootstrap method.

3.2 Variance and Covariance estimates using the bootstrap method

The nonparametric bootstrap method (Efron and Tibshirani, 1993) was conducted by randomly selecting subjects with replacement from the original data. For instance, for the UTI data, we resample N_k women with replacement from the k^{th} stratum, where $k = 1, 2$. Similarly, for the Linguistics example, we resample 50 utterances with replacement and cross classify the data into a $3 \times 7 \times 6$ table. For each resampled data set, the size of each stratum is the same as the original data. We take B resamples and then for each resample, we calculate the generalized MH estimates $\{\bar{L}^x_{ab},\, x = 1, \ldots, m,\, a \neq b = 1, \ldots, r\}$. The bootstrap estimate of the standard error of $\bar{L}^x_{ab}$ is the standard deviation of the bootstrap replicates,

$$\text{s.e. for } \bar{L}^x_{ab} = \sqrt{\frac{\sum_{s=1}^{B}\left(\bar{L}^x_{ab,s} - \sum_{s=1}^{B}\bar{L}^x_{ab,s}/B\right)^2}{B-1}},$$

where $\bar{L}^x_{ab,s}$ is the generalized MH estimate $\bar{L}^x_{ab}$ for the s^{th} bootstrap resample. Similarly, the bootstrap estimate of the covariance of $\bar{L}^x_{ab}$ and $\bar{L}^y_{cd}$ is

$$\hat{\mathrm{cov}}(\bar{L}^x_{ab},\, \bar{L}^y_{cd}) = \frac{\sum_{s=1}^{B}\left(\bar{L}^x_{ab,s} - \sum_{s=1}^{B}\bar{L}^x_{ab,s}/B\right)\left(\bar{L}^y_{cd,s} - \sum_{s=1}^{B}\bar{L}^y_{cd,s}/B\right)}{B-1}.$$

Later, we will refer to this as "bootstrap" estimate.

4 Examples

For the UTI example, the model-based (GEE) approach gives $\{\hat{\beta}_{1j},\, j = 1, \ldots, 5\} = \{0.12\ ,\ -0.52,\ 0.71,\ 0.65,\ -8.96\}$ with sandwich standard errors $\{0.28, 0.27, 0.28, 0.31, 1.10\}$ using an

exchangeable correlation structure. Alternatively, the non-model-based (MH) approach gives $\{L_{12}^j,\ j=1,\ldots,5\}$=$\{0.12\ ,\ -0.52,\ 0.71,\ 0.64,\ -2.57\}$ with standard errors $\{0.28,\ 0.26,\ 0.28,\ 0.31,\ 1.41\}$ by applying formula (2). Choosing $B=100$, the corresponding bootstrap standard error is $\{0.28,\ 0.26,\ 0.28,\ 0.32,\ 0.39\}$. For instance, for the first item (oral contraceptive), the odds of having used the oral contraceptive for women without a prior history of UTI are estimated to be $\exp(0.12)=1.13$ times higher than the odds for women with a prior history of UTI, given each age group.

Table 4: The "bootstrap" and "formulae" (in parentheses) variance and covariance estimates of $\{L_{12}^j,\ j=1,\ldots,5\}$,

	L_{12}^1	L_{12}^2	L_{12}^3	L_{12}^4	L_{12}^5
L_{12}^1	$0.079(0.076)$	$-0.050(-0.048)$	$-0.045(-0.042)$	$-0.048(-0.002)$	$0.011(NA)$
L_{12}^2	$-0.050(-0.048)$	$0.068(0.070)$	$0.051(-0.037)$	$0.045(0.060)$	$-0.007(NA)$
L_{12}^3	$-0.045(-0.042)$	$0.051(-0.037)$	$0.081(0.080)$	$0.051(0.044)$	$-0.006(NA)$
L_{12}^4	$-0.048(-0.002)$	$0.045(0.060)$	$0.051(0.044)$	$0.104(0.094)$	$-0.012(NA)$
L_{12}^5	$0.011(NA)$	$-0.007(NA)$	$-0.006(NA)$	$-0.012(NA)$	$0.152(1.994)$

NA: not applicable

Table 5: A 95% confidence interval for $\log\Psi_{12}^x - \log\Psi_{12}^y$.

	x				
	1	2	3	4	5
y	Oral	Condom	L. cond.	Spermicide	Diaphragm
1. Oral		$(-1.6140,$ $0.3342)$	$(-0.3876,$ $1.5724)$	$(-0.5117,$ $1.5589)$	$(-3.5851,$ $-1.7931)$
2. Condom	$(-1.6055,$ $0.3257)$		$(0.8074,$ $1.6572)$	$(0.6022,$ $1.7248)$	$(-2.9973,$ $-1.1011)$
3. L. cond.	$(-0.3684,$ $1.5532)$	$(0.3034,$ $2.1613)$		$(-0.6335,$ $0.4959)$	$(-4.2517,$ $-2.3113)$
4. Spermicide	(NA)	(NA)	(NA)		$(-4.2498,$ $-2.1756)$

based on formulae (lower left half) and bootstrap (upper right half) (co)variance estimates
NA: not applicable

The two approaches have similar results, except for the last item ("Diaphragm"), because our data have no women without prior history of urinary tract infection who use diaphragms. In Table 1, the cell count for row 1 and column 5 is zero for both age groups. For GEE approach, the estimation routine fails to provide sandwich standard errors. Similarly, MH estimate L_{12}^5 is undefined. To fix the problem for the model-based approach, we add a pseudo subject who didn't have UTI history but used Diaphragm into the dataset. The model (1)

Table 6: The generalized MH estimates and their bootstrap standard errors (in parentheses) for the data in table 2

	item j						
	1 pronunciation of consonants	2 pronunciation of vowels	3 word stress	4 sentence stress	5 rhythm	6 intonation	7 rate
$\bar{L}_{12}^j$	−0.00	1.19	0.70	0.28	−0.10	0.88	−0.39
	(0.81)	(0.50)	(0.53)	(0.40)	(0.47)	(0.50)	(1.07)
$\bar{L}_{13}^j$	1.34	1.47	1.21	1.49	0.73	1.36	−1.23
	(0.73)	(0.48)	(0.58)	(0.49)	(0.44)	(0.45)	(1.17)
$\bar{L}_{23}^j$	1.34	0.27	0.52	1.20	0.83	0.48	−0.84
	(0.52)	(0.30)	(0.47)	(0.50)	(0.50)	(0.43)	(1.35)

is fitted by giving the pseudo subject a small weight (say, 10^{-3}). For the non-model-based approach, one way to get an amended estimator is by adding 0.5 to each cell as suggested by Agresti (2002, p. 71) for the ordinary odds ratio estimator. The cell counts in the stratum having few observations are usually small. If we add 0.5 to a small cell count, it could easily influence the association which weakens the association. In order not to smooth the data too much, we add 0.5 to each cell for the stratum with largest size. For instance, because the stratum of Age<24 contains the greater number of observations, we add 0.5 to each cell in that stratum. The estimate of the odds ratio for the last item ("Diaphragm") is not stable for two approaches because of the imputation. In summary, the conditional UTI effects are significant for the contraceptives "condom", "lubricated condom", and "spermicide" at a 5% significance level.

Table 4 gives the bootstrap and formulae (in parentheses) variance and covariance estimates for $\{L_{12}^j,\, j = 1, \ldots, 5\}$. Table 5 shows all multiple comparisons of the conditional UTI effects for any two items. For instance, comparing the UTI effects for the contraceptives "oral" and "lubricated condom", a 95% confidence interval for $\log \Psi_{12}^3 - \log \Psi_{12}^1$ gives (−0.39, 1.57). Due to the sampling zero for the item 5 (Diaphragm), a few consistent covariance estimators involving L_{12}^5 are not applicable. Consequently, the confidence intervals based on the formulae are not applicable for item 5. Alternatively, one can choose to amend the pairwise observations to obtain rough estimates for them.

For the Linguistics example, the GEE approach fails to give the sandwich standard errors for the model (1). Alternatively, we fit a parsimonious model that replaces τ_{jk} by $\tau_j + \alpha_k$. However, the generalized MH estimator works for the general model (1). By comparing overall rating levels 1 and 2, the MH estimates $\{\bar{L}_{12}^j,\, j = 1, \ldots, 7\}$ are {−0.00, 1.19, 0.70, 0.28, −0.10, 0.88, −0.39} with the bootstrap standard error of {0.81, 0.50, 0.53, 0.40, 0.47, 0.50, 1.07}. Comparing rating levels 1 and 3, the MH estimates $\{\bar{L}_{13}^j,\, j = 1, \ldots, 7\}$ is {1.34, 1.47, 1.21, 1.49, 0.73, 1.36, −1.23} with the bootstrap standard error of {0.73, 0.48, 0.58, 0.49, 0.44, 0.45, 1.17}. There are no significant differences between rating levels 1 and 2 for most of items, except for item 2 (pronunciation of vowels), given each of raters. However, the differences between rating levels 1 and 3 are significant for most of items given each rater, except for items 1, 5, and 7. Table 6 shows the generalized MH estimates and their bootstrap standard errors. Similarly, the bootstrap variances and covariances estimates can be calculated. For this example, the formulae (co)variance estimators are not appropriate, because this dataset

has dependent strata.

Although the GEE results are not comparable because it uses a more parsimonious model, they give similar results in terms of the significance. For instance, the GEE estimates for $\{\log \Psi_{13}^j, j = 1, \ldots, 7\}$ are $\{1.34,\ 1.32\,0.83,\ 1.29,\ 0.76,\ 1.24,\ -1.11\}$ with the sandwich standard errors $\{0.67,\ 0.39,\ 0.53,\ 0.38,\ 0.33,\ 0.32,\ 1.19\}$.

5 Simulation Study

In the simulation study we evaluate the performance of the model based (GEE) and non-model based (MH) estimators for the odds ratio and their (co)variances estimators. The simulation study consists of two main cases. One case assumes that the strata are independent as in the UTI example. The other case allows dependency between strata as in the Linguistics example. For case 1, the scenarios range from ones for which the asymptotic type (a) should work well to ones for which the asymptotic type (b) seems more appropriate. For case 2, the situations vary based on the degree of the dependency between strata.

For the model based estimators (GEE), we use R (R Development Core Team) and its package "geepack" (Yan, 2002; Yan and Fine, 2004) for fitting. We always assume an exchangeable correlation structure to obtain the estimates $\{\hat{\gamma}_{ab}^j; a \neq b; a, b = 1, \ldots, r; j = 1, \ldots, m\}$. We automatically yield the robust (or sandwich) and naive (co)variances as a by-product from the fitting algorithm. For the non-model based method (MH) we compute $\{\bar{L}_{ab}^j;\ a \neq b; a, b = 1, \ldots, r; j = 1, \ldots, m\}$ and its bootstrap and formulae (co)variances.

Independent Strata For simplicity we let $r = 2$ and use a constant odds ratio for every item, i.e., $\Psi_{12}^j = \Psi$ for all $j = 1, \ldots, m$. We also set the marginal probabilities $\pi_{j|1k}$ to be 0.5 for all items $j = 1, \ldots, m$ and strata $k = 1, \ldots, K$. The marginal probabilities $\{\pi_{j|2k}\}$ are computed from the given common odds ratio Ψ. Let Y_j denote whether a subject selects item j. Given i and k, if a subject selects item j, then $Y_j = 1$; otherwise, $Y_j = 0$. The pairwise dependency between items x and y is denoted using an odds ratio θ_{xy} as

$$\theta_{xy} = \frac{P(Y_x = 1, Y_y = 1)P(Y_x = 0, Y_y = 0)}{P(Y_x = 0, Y_y = 1)P(Y_x = 1, Y_y = 0)},$$

where $x \neq y = 1, \ldots, m$. Then, the 2^m joint probabilities $P_{Y|ik} = \{P(Y_1 = b_1, \ldots, Y_m = b_m|ik), b_j = 0, 1; j = 1, \ldots, m\}$ in the complete table as in Table 3 can be computed from the marginal probabilities $\{\pi_{j|ik}\}$ and $\{\theta_{xy}, x \neq y = 1, \ldots, m\}$ described by Lee (1993) and by Gange (1995) if a feasible solution exists. Following the simulation scheme by Bilder, Loughin, and Nettleton (2000), we apply Gange's method, because it generates strictly positive (> 0) joint probabilities in most cases. It seems more plausible, since none of the 2^m binary sequences is theoretically excluded from the data generation process.

Again, for simplicity, we let $m = 2$. The dependency between items is assigned by the odds ratio $\theta = \theta_{12}$. We draw N_k samples independently from either row 1 or row 2 with equal probabilities for stratum k and set $N_1 = \ldots = N_K$. Given the randomly chosen row i and stratum k, a sample (a binary sequence of length m), is drawn from the joint distribution $P_{Y|ik}$. In case 1, we simulate $n = 20000$ datasets based on the joint distributions $\{P_{Y|ik}\}$ under a variety of configurations. For the bootstrap method, we use the number of bootstrap resamples $B = 400$. Note that, we did not compute the model nor non-model based estimators, when data amendment was required due to the sampling zero problem.

Table 7 shows the sample means for the generalized MH estimates (L_{12}^1, L_{12}^2) in the first row, and the sample means for the GEE estimates ($\hat{\gamma}_{12}^1, \hat{\gamma}_{12}^2$) in the second row over $n = 20000$ simulations for various scenarios. We investigate (1) the performance of the MH (L's) and the GEE ($\hat{\gamma}$'s) by comparing their sample means and the mean square errors (mse), (2) the performance of the MH (co)variance estimators for the formulae and bootstrap methods, and (3) the performance of GEE (co)variance estimators for the robust and naive methods. To compare the performance of the (co)variance estimators, we calculate the "empirical" (co)variances. For instance, the empirical (co)variance of L_{12}^1 is defined as the sample (co)variance of L_{12}^1 over 20000 simulations.

For the non-model based approach, we denote the sample mean for formulae (co)variances by formulae$_{MH}$, and the bootstrap (co)variances by bootstrap$_{MH}$. The empirical (co)variance is denoted by empirical$_{MH}$. Similarly we denote for the model based approach the empirical (co)variances by empirical$_{GEE}$, the mean of the robust and naive (co)variances by robust$_{GEE}$ and naive$_{GEE}$, respectively. Each entry in Table 7 consists of three terms. The first two are the variances of the log odds ratio (L's or $\hat{\gamma}$'s) for items 1 and 2, and the third is the covariance of the log odds ratios between items 1 and 2. The first column shows the configuration of parameters K, N_k, Ψ, θ, and the number in parentheses shows the number of samples which were not included in the simulation study due to the sampling zero problem. The total number of simulated samples involved is 20000 − (this number).

Dependent Strata In case 2, we let $r = m = 2$. Unlike case 1, there is some degree of dependency between strata (or raters in the Linguistics example). We introduce another two parameters Λ_{uv} and $\Gamma_{xy,uv}$ to describe the dependencies between items and between raters. Let Z_k be whether rater k assigns an overall rating 1. If it is a "yes", then $Z_k = 1$; otherwise $Z_k = 0$. Similarly, let $W_{j,k}$ be whether rater k selects item j. If rater k selects item j, then $W_{j,k} = 1$; otherwise $W_{j,k} = 0$. The parameters Λ_{uv} and $\Gamma_{xy,uv}$ are defined as

$$\begin{aligned}
\Lambda_{uv} &= \frac{P(Z_u = 1, Z_v = 1)P(Z_u = 0, Z_v = 0)}{P(Z_u = 0, Z_v = 1)P(Z_u = 1, Z_v = 0)}, \quad u \neq v = 1, \ldots, K; \\
\Gamma_{xy,uv} &= \frac{P(W_{x,u} = 1, W_{y,v} = 1)P(W_{x,u} = 0, W_{y,v} = 0)}{P(W_{x,u} = 0, W_{y,v} = 1)P(W_{x,u} = 1, W_{y,v} = 0)}, \\
&\quad u \neq v = 1, \ldots, K \text{ or } x \neq y = 1, \ldots, m.
\end{aligned}$$

For a special case of $u = v$, $\Gamma_{xy,uv} = \theta_{xy}$ describes the dependency between items for a given rater k. In contrast, $\Gamma_{xy,uv}$ with $u \neq v$ denotes the dependency between items and between raters. For convenience, we set $\Lambda_{uv} = \Lambda$ for all $u < v = 1, \ldots, K$; $\Gamma_{12,kk} = \theta$ for all $k = 1, \ldots, K$; and $\Gamma_{xy,uv} = \Gamma$ for all $u < v = 1, \ldots, K$ and $x \leq y = 1, 2$.

We first fix the marginal overall rating probabilities $P(Z_k = 1) = 0.5, k = 1, \ldots, K$ and compute the overall rating joint probabilities $P_Z = \{P(Z_1 = z_1, \ldots, Z_K = z_K), z_k = 0, 1; k = 1, \ldots, K\}$ from $\{P(Z_k = 1)\}$ and Λ applying Gange's method. As in case 1, $\pi_{j|1k}$ is set to be 0.5 for all items $j = 1, \ldots, m$ and strata $k = 1, \ldots, K$. The marginal probabilities $\{\pi_{j|2k}\}$ are computed from the given common odds ratio Ψ. Given a specific overall rating configuration $z = (z_1, \ldots, z_K)$, the items joint distribution $P_{W|z} = \{P(W_{1,1} = w_{1,1}, \ldots, W_{m,1} = w_{m,1}, \ldots, W_{1,K} = w_{1,K}, \ldots, W_{m,K} = w_{m,K}|z)$,$w_{j,k} = 0, 1; j = 1, \ldots, m; k = 1, \ldots, K\}$ can be computed from $\{\pi_{j|ik}\}$, θ and Γ using Gange's method. The 2^K possible overall ratings configurations result in 2^K different items joint distributions $P_{W|z}$, which are all computed in advance.

Then, we draw $N_k = N$ samples from the overall rating joint distribution P_Z. Now given such a realization z, we can sample one vector of length $m \cdot K$ from $P_{W|z}$. Then, we separate

Table 7: MH and GEE Results of the simulation study for independent strata

K, N_k, Ψ, θ	mean $(L_{12}^1, L_{12}^1)_{MH}$ $(\hat{\gamma}_{12}^1, \hat{\gamma}_{12}^2)_{GEE}$	$^{10\cdot}\text{mse}_{MH}$ $^{10\cdot}\text{mse}_{GEE}$	$\text{Var}(L/\hat{\gamma})_{12}^1, \text{Var}(L/\hat{\gamma})_{12}^2, \text{Cov}\{(L/\hat{\gamma})_{12}^1, (L/\hat{\gamma})_{12}^2\}$ $^{10\cdot}\text{empirical}_{MH}$ $^{10\cdot}\text{empirical}_{GEE}$	$^{10\cdot}\text{formulae}_{MH}$ $^{10\cdot}\text{robust}_{GEE}$	$^{10\cdot}\text{bootstrap}_{MH}$ $^{10\cdot}\text{naive}_{GEE}$
$2, 50, 1, 2$	$-0.000, 0.002$	$1.75, 1.70$	$1.75, 1.70, 0.310$	$1.67, 1.67, 0.279$	$1.77, 1.77, 0.300$
$^{(4)}$	$-0.000, 0.002$	$1.78, 1.73$	$1.78, 1.73, 0.317$	$1.71, 1.71, 0.291$	$1.71, 1.71, 0.291$
$2, 50, 1, 4$	$-0.001, 0.002$	$1.75, 1.71$	$1.75, 1.71, 0.588$	$1.67, 1.67, 0.543$	$1.77, 1.77, 0.584$
$^{(3)}$	$-0.001, 0.002$	$1.78, 1.75$	$1.78, 1.75, 0.601$	$1.71, 1.71, 0.566$	$1.71, 1.71, 0.566$
$^{(2)}2, 50, 4, 2$	$1.425, 1.428$	$2.33, 2.32$	$2.32, 2.30, 0.296$	$2.23, 2.23, 0.294$	$2.55, 2.55, 0.331$
$^{(2)}$	$1.440, 1.443$	$2.39, 2.38$	$2.36, 2.34, 0.304$	$2.27, 2.27, 0.308$	$2.27, 2.27, 0.323$
$^{(1)}2, 50, 4, 4$	$1.429, 1.434$	$2.33, 2.36$	$2.32, 2.34, 0.654$	$2.24, 2.24, 0.611$	$2.56, 2.56, 0.687$
$^{(1)}$	$1.443, 1.450$	$2.39, 2.43$	$2.36, 2.39, 0.668$	$2.27, 2.28, 0.638$	$2.28, 2.28, 0.660$
$2, 100, 1, 2$	$-0.002, -0.002$	$0.84, 0.83$	$0.84, 0.83, 0.148$	$0.82, 0.82, 0.138$	$0.84, 0.84, 0.142$
$^{(3)}$	$-0.002, -0.002$	$0.85, 0.84$	$0.85, 0.84, 0.149$	$0.83, 0.83, 0.141$	$0.83, 0.83, 0.141$
$2, 100, 1, 4$	$-0.002, -0.003$	$0.84, 0.83$	$0.84, 0.83, 0.280$	$0.82, 0.82, 0.269$	$0.84, 0.84, 0.278$
$^{(2)}$	$-0.002, -0.003$	$0.85, 0.84$	$0.85, 0.84, 0.283$	$0.83, 0.83, 0.274$	$0.83, 0.83, 0.275$
$2, 100, 4, 2$	$1.408, 1.405$	$1.08, 1.09$	$1.07, 1.09, 0.144$	$1.07, 1.06, 0.148$	$1.13, 1.13, 0.156$
	$1.415, 1.412$	$1.09, 1.10$	$1.09, 1.10, 0.146$	$1.07, 1.07, 0.151$	$1.08, 1.07, 0.157$
$2, 100, 4, 4$	$1.407, 1.405$	$1.08, 1.09$	$1.08, 1.08, 0.309$	$1.06, 1.06, 0.301$	$1.13, 1.13, 0.317$
	$1.414, 1.412$	$1.10, 1.10$	$1.09, 1.09, 0.313$	$1.07, 1.07, 0.307$	$1.07, 1.07, 0.316$
$20, 5, 1, 2$	$-0.002, 0.005$	$2.25, 2.22$	$2.25, 2.22, 0.382$	$2.12, 2.12, 0.351$	$2.30, 2.30, 0.362$
$^{(17653)}$	$-0.007, -0.005$	$2.83, 2.89$	$2.83, 2.89, 0.527$	$2.47, 2.48, 0.371$	$2.45, 2.45, 0.358$
$20, 5, 1, 4$	$0.003, -0.003$	$2.21, 2.20$	$2.21, 2.20, 0.739$	$2.12, 2.12, 0.678$	$2.29, 2.29, 0.709$
$^{(18027)}$	$0.003, 0.014$	$3.19, 3.00$	$3.19, 3.00, 1.150$	$2.46, 2.46, 0.838$	$2.45, 2.45, 0.824$
$^{(47)}20, 5, 4, 2$	$1.464, 1.460$	$3.56, 3.48$	$3.50, 3.43, 0.391$	$3.22, 3.20, 0.371$	$3.50, 3.50, 0.354$
$^{(19751)}$	$1.890, 1.779$	$7.95, 6.75$	$5.44, 5.22, 0.684$	$3.65, 3.48, 0.319$	$3.55, 3.40, 0.379$
$^{(28)}20, 5, 4, 4$	$1.457, 1.463$	$3.47, 3.47$	$3.42, 3.41, 0.881$	$3.19, 3.20, 0.769$	$3.48, 3.48, 0.743$
$^{(19784)}$	$1.916, 1.945$	$7.78, 9.16$	$4.99, 6.07, 1.565$	$3.68, 3.75, 1.057$	$3.63, 3.69, 1.110$
$20, 10, 1, 2$	$-0.000, -0.003$	$0.93, 0.92$	$0.93, 0.92, 0.162$	$0.91, 0.90, 0.152$	$0.88, 0.88, 0.148$
$^{(1116)}$	$0.000, -0.004$	$1.13, 1.11$	$1.13, 1.11, 0.202$	$1.00, 1.00, 0.170$	$1.00, 1.00, 0.168$
$20, 10, 1, 4$	$-0.001, 0.001$	$0.92, 0.93$	$0.92, 0.93, 0.313$	$0.90, 0.91, 0.298$	$0.88, 0.88, 0.290$
$^{(1227)}$	$-0.002, 0.000$	$1.12, 1.12$	$1.12, 1.12, 0.386$	$1.00, 1.00, 0.334$	$1.00, 1.00, 0.331$
$20, 10, 4, 2$	$1.411, 1.412$	$1.27, 1.28$	$1.27, 1.28, 0.170$	$1.23, 1.24, 0.162$	$1.38, 1.39, 0.167$
$^{(7085)}$	$1.569, 1.570$	$1.90, 1.87$	$1.56, 1.54, 0.219$	$1.35, 1.34, 0.181$	$1.34, 1.34, 0.184$
$20, 10, 4, 4$	$1.412, 1.414$	$1.28, 1.28$	$1.27, 1.27, 0.346$	$1.23, 1.24, 0.335$	$1.39, 1.39, 0.349$
$^{(7667)}$	$1.571, 1.572$	$1.86, 1.91$	$1.52, 1.56, 0.428$	$1.34, 1.34, 0.382$	$1.34, 1.34, 0.383$

$\log(1) = 0, \log(4) = 1.3863$. $n = 20000$ and $B = 400$.
The value in parentheses is number of datasets having the sampling zero problem (not included)

each of the vectors of length mK in K vectors of length m, such that the k^{th} vector of length m represents the items of rater k. For instance, for $m = 2$, if the k^{th} vector is (0, 1), then it says that rater k selects item 2, but not item 1. We draw samples from P_Z in order to incorporate some dependency in the overall rating between raters.

In case 2, it is not feasible to sample sparse data with a large number of strata ($K >> 5$). Choosing $K = 5$ and $m = 2$, we already get $2^{mK} = 2^{10} = 1024$ joint probabilities in $P_{W|z}$ for each overall ratings configuration z. Increasing m or K creates a problem with a huge number of joint probabilities which is infeasible for most computers. In total, we simulate $n = 20000$ datasets under a variety of configurations. For the bootstrap method, we use the number of bootstrap resamples as $B = 400$. Table 8 shows the results using the same notation as Table 7.

Results Table 7 shows that the MH approach performs better than the GEE approach, especially when N_k is small. Also, GEE often fails to converge for extremely sparse data, e.g., $N_k = 5$. The convergence problem occurs when the number of parameters increases as the number of strata increases. In contrast, table 8 shows that GEE provides better estimates for high dependence ($\Gamma \geq 4$) between strata, whereas for low dependence ($\Gamma = 2$) MH still performs similarly well as GEE.

When we compare the bootstrap with the formulae (co)variances, we can say the following. Under independence of strata the formulae (co)variance and bootstrap (co)variance behave similarly. For the dependent strata case, the bootstrap (co)variance is better than the formulae (co)variance. Only for a few configurations ($\Gamma = 2$) the formulae (co)variance is still quite good and similar to the bootstrap (co)variance despite the violation of the naive independence assumption.

Comparing the (co)variance estimates for GEE, we see that the robust (co)variance is generally better than the naive as expected, because the naive (co)variance assumes that the correlation structure chosen is the correct one. In case 1, the dependence only occurs across 2 different items. Since the deviation of the chosen correlation structure from the empirical one is not severe, the naive and robust (co)variances perform quite similarly. For case 2, dependence occurs across different items and strata. The performance of the naive (co)variance becomes poorly.

Most software like R only offer simple choices such as "exchangeable", "unstructured", "independence" for all observations and ratings, and one cannot match the exact correlation structure as in our simulation study. The "exchangeable" structure is the most common one, because it incorporates fewer parameters that result in less convergence problems.

6 Conclusion

This article uses both the model-based (GEE) and non-model-based (MH) approaches to evaluate the conditional associations between row and column variables for each of the items for stratified multiple responses. The model-based approach is suitable if one is interested in the model selection in order to find the relationship between the item responses and explanatory variables. For highly sparse data (K large, but N_k small), it might have convergence problems. However, if one is particularly interested in the conditional association between the item and the explanatory variable given strata, the MH-type estimators evaluate the association directly. From the simulation studies, they agree with each other.

We give two examples in this paper. The UTI example has independent strata and the Linguistic example has dependent strata. For the MH approach with independent strata, Greenland (1989) provided dually consistent variance and covariance estimators for single items, whereas we derived dually consistent covariance estimators between items. For depen-

Table 8: MH and GEE Results of the simulation study for dependent strata

K, N_k, Λ, Γ	mean $(L^1_{12}, L^1_{12})_{MH}$ $(\hat{\gamma}^1_{12}, \hat{\gamma}^2_{12})_{GEE}$	$^{10\cdot}\text{mse}_{MH}$ $^{10\cdot}\text{mse}_{GEE}$	$\text{Var}(L/\hat{\gamma})^1_{12}, \text{Var}(L/\hat{\gamma})^2_{12}, \text{Cov}\{(L/\hat{\gamma})^1_{12}, (L/\hat{\gamma})^2_{12}\}$ $^{10\cdot}\text{empirical}_{MH}$ $^{10\cdot}\text{empirical}_{GEE}$	$^{10\cdot}\text{formulae}_{MH}$ $^{10\cdot}\text{robust}_{GEE}$	$^{10\cdot}\text{bootstrap}_{MH}$ $^{10\cdot}\text{naive}_{GEE}$
$^{(1)}2, 50, 2, 2$	$1.434, 1.431$	$2.41, 2.40$	$2.39, 2.38, 0.728$	$2.24, 2.24, 0.607$	$2.64, 2.63, 0.749$
$^{(1)}$	$1.446, 1.442$	$2.44, 2.43$	$2.40, 2.40, 0.717$	$2.27, 2.27, 0.628$	$2.23, 2.22, 0.358$
$^{(4)}2, 50, 2, 4$	$1.427, 1.432$	$2.57, 2.55$	$2.56, 2.53, 0.830$	$2.24, 2.25, 0.605$	$2.73, 2.74, 0.835$
$^{(4)}$	$1.439, 1.445$	$2.36, 2.37$	$2.33, 2.33, 0.600$	$2.15, 2.16, 0.521$	$2.11, 2.12, 0.454$
$^{(4)}2, 50, 4, 2$	$1.434, 1.426$	$2.48, 2.46$	$2.46, 2.44, 0.779$	$2.24, 2.24, 0.607$	$2.70, 2.68, 0.802$
$^{(4)}$	$1.448, 1.440$	$2.53, 2.48$	$2.49, 2.45, 0.779$	$2.33, 2.32, 0.681$	$2.30, 2.29, 0.432$
$^{(2)}2, 50, 4, 4$	$1.429, 1.431$	$2.66, 2.60$	$2.65, 2.58, 0.941$	$2.26, 2.25, 0.597$	$2.84, 2.84, 0.925$
$^{(2)}$	$1.441, 1.444$	$2.45, 2.41$	$2.42, 2.37, 0.714$	$2.26, 2.26, 0.607$	$2.22, 2.22, 0.557$
$^{(2)}2, 50, 4, 9$	$1.443, 1.443$	$2.86, 2.94$	$2.83, 2.90, 1.148$	$2.28, 2.28, 0.591$	$3.04, 3.05, 1.105$
$^{(2)}$	$1.449, 1.448$	$2.34, 2.36$	$2.30, 2.32, 0.595$	$2.12, 2.12, 0.469$	$2.09, 2.09, 0.640$
$2, 100, 2, 2$	$1.411, 1.411$	$1.13, 1.13$	$1.13, 1.12, 0.347$	$1.07, 1.07, 0.301$	$1.17, 1.17, 0.349$
	$1.417, 1.418$	$1.12, 1.13$	$1.12, 1.12, 0.334$	$1.08, 1.08, 0.310$	$1.05, 1.05, 0.174$
$2, 100, 2, 4$	$1.406, 1.413$	$1.17, 1.15$	$1.17, 1.15, 0.376$	$1.07, 1.07, 0.299$	$1.21, 1.21, 0.389$
	$1.412, 1.419$	$1.07, 1.05$	$1.07, 1.04, 0.270$	$1.02, 1.02, 0.252$	$0.99, 1.00, 0.219$
$2, 100, 2, 9$	$1.411, 1.409$	$1.22, 1.25$	$1.22, 1.25, 0.451$	$1.07, 1.07, 0.298$	$1.28, 1.28, 0.446$
	$1.414, 1.412$	$0.96, 0.98$	$0.96, 0.97, 0.183$	$0.93, 0.93, 0.164$	$0.92, 0.92, 0.243$
$2, 100, 4, 2$	$1.408, 1.410$	$1.15, 1.13$	$1.15, 1.13, 0.359$	$1.07, 1.07, 0.301$	$1.19, 1.19, 0.373$
	$1.415, 1.417$	$1.15, 1.15$	$1.14, 1.14, 0.354$	$1.10, 1.10, 0.336$	$1.08, 1.08, 0.210$
$2, 100, 4, 4$	$1.411, 1.407$	$1.21, 1.20$	$1.21, 1.19, 0.429$	$1.07, 1.07, 0.298$	$1.26, 1.26, 0.435$
	$1.416, 1.412$	$1.12, 1.09$	$1.11, 1.09, 0.320$	$1.07, 1.07, 0.299$	$1.04, 1.04, 0.269$
$^{(3)}5, 20, 1, 2$	$1.432, 1.436$	$2.63, 2.57$	$2.61, 2.55, 0.760$	$2.38, 2.38, 0.629$	$2.94, 2.95, 0.762$
$^{(38)}$	$1.489, 1.492$	$2.73, 2.69$	$2.62, 2.58, 0.665$	$2.32, 2.33, 0.545$	$2.37, 2.36, 0.172$
$^{(5)}5, 20, 2, 2$	$1.444, 1.441$	$2.88, 2.79$	$2.84, 2.76, 0.990$	$2.40, 2.39, 0.620$	$3.18, 3.19, 0.978$
$^{(34)}$	$1.500, 1.499$	$2.91, 2.94$	$2.78, 2.81, 0.830$	$2.46, 2.47, 0.663$	$2.50, 2.51, 0.302$
$^{(10)}5, 20, 2, 4$	$1.455, 1.454$	$3.21, 3.25$	$3.16, 3.20, 1.310$	$2.43, 2.43, 0.597$	$3.54, 3.53, 1.256$
$^{(46)}$	$1.506, 1.504$	$2.75, 2.77$	$2.61, 2.63, 0.670$	$2.34, 2.35, 0.535$	$2.28, 2.29, 0.368$
$^{(32)}5, 20, 2, 9$	$1.458, 1.462$	$3.93, 3.83$	$3.87, 3.78, 1.852$	$2.50, 2.50, 0.565$	$4.02, 4.03, 1.705$
$^{(66)}$	$1.497, 1.500$	$2.78, 2.69$	$2.66, 2.56, 0.628$	$2.30, 2.30, 0.488$	$2.02, 2.02, 0.394$
$^{(8)}5, 20, 4, 2$	$1.442, 1.445$	$3.07, 3.12$	$3.03, 3.09, 1.203$	$2.40, 2.41, 0.617$	$3.37, 3.40, 1.179$
$^{(41)}$	$1.500, 1.503$	$3.09, 3.14$	$2.96, 3.01, 1.016$	$2.59, 2.60, 0.805$	$2.67, 2.67, 0.454$
$^{(21)}5, 20, 4, 4$	$1.464, 1.459$	$3.68, 3.71$	$3.62, 3.66, 1.734$	$2.47, 2.46, 0.582$	$3.97, 3.97, 1.664$
$^{(53)}$	$1.505, 1.500$	$3.02, 3.03$	$2.88, 2.90, 0.945$	$2.54, 2.54, 0.734$	$2.49, 2.48, 0.568$
$^{(48)}5, 20, 4, 9$	$1.467, 1.463$	$4.55, 4.42$	$4.49, 4.36, 2.447$	$2.55, 2.53, 0.549$	$4.67, 4.67, 2.317$
$^{(91)}$	$1.500, 1.497$	$2.99, 2.93$	$2.86, 2.80, 0.875$	$2.53, 2.52, 0.701$	$2.22, 2.19, 0.596$
$5, 100, 2, 2$	$1.396, 1.392$	$0.48, 0.47$	$0.48, 0.47, 0.175$	$0.42, 0.42, 0.120$	$0.48, 0.48, 0.169$
	$1.407, 1.404$	$0.44, 0.44$	$0.44, 0.43, 0.131$	$0.43, 0.43, 0.124$	$0.41, 0.41, 0.054$
$5, 100, 2, 4$	$1.397, 1.393$	$0.54, 0.54$	$0.54, 0.53, 0.233$	$0.42, 0.42, 0.119$	$0.54, 0.54, 0.231$
	$1.407, 1.403$	$0.42, 0.41$	$0.41, 0.41, 0.103$	$0.40, 0.40, 0.101$	$0.37, 0.37, 0.063$
$5, 100, 4, 2$	$1.396, 1.391$	$0.52, 0.51$	$0.52, 0.51, 0.213$	$0.42, 0.42, 0.119$	$0.51, 0.51, 0.205$
	$1.407, 1.402$	$0.47, 0.46$	$0.47, 0.46, 0.160$	$0.45, 0.45, 0.150$	$0.44, 0.44, 0.081$
$5, 100, 4, 4$	$1.396, 1.397$	$0.61, 0.62$	$0.61, 0.61, 0.305$	$0.42, 0.42, 0.118$	$0.62, 0.62, 0.305$
	$1.405, 1.406$	$0.45, 0.46$	$0.45, 0.46, 0.143$	$0.44, 0.44, 0.137$	$0.40, 0.40, 0.097$

$\Psi = 4$ and $\theta = 4$ for all configurations. $n = 20000$ and $B = 400$.
The value in parentheses is number of datasets having the sampling zero problem (not included)

dent data the bootstrap method provides an easy and plausible way to estimate variances and covariances. It also performs as well as the formulae estimates for the independent strata cases.

The Linguistic example is a case of multilevel data where there is a hierarchical correlated structure to the data. The responses are correlated within each of the m items; and within each item, the responses are correlated within each of the K raters. Besides the GEE and MH methods, a generalized linear mixed model can also be used for analyzing the multilevel data. Fitzmaurice, Laird, and Ware (2004) discuss the multilevel generalized linear mixed model. Unfortunately, using the existing software it is not easy to implement the multilevel generalized linear mixed model. Users need to write their own programs for this.

Acknowledgement

We would like to thank Megan Clark for helpful comments.

References

[1] Agresti, A. (2002). *Categorical Data Analysis, 2nd edition.* New Jersey: Wiley.

[2] Agresti, A. and Liu, I. (1999). Marginal modeling of a categorical variable allowing arbitrarily many category choices. *Biometrics* 55:936-43.

[3] Agresti, A. and Liu, I. (2001). Strategies for modeling a categorical variable allowing multiple category choices. *Sociol. Methods Res.* 29:403-434.

[4] Andersen, E. B. (1980). *Discrete statistical methods with social science applications.* New York: North-Holland.

[5] Bilder, C. R., Loughin, T. M., and Nettleton, D. (2000). Multiple marginal independence testing for pick any/c variables. *Commun. Statist. – Comput. Simu.* 29:1285-1316.

[6] Bilder, C. R. and Loughin, T. M. (2002). Testing for conditional multiple marginal independence. *Biometrics* 58:200-208.

[7] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* New York: Chapman and Hall.

[8] Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004). *Applied Longitudinal Analysis.* New Jersey: Wiley.

[9] Gange, S. J. (1995). Generating multivariate categorical variates using the iterative proportional fitting algorithm. *Amer. Statist.* 49:134-138.

[10] Greenland, S. (1989). Generalized Mantel-Haenszel estimators for K $2 \times J$ tables. *Biometrics* 45:183-191

[11] Lee, A. J. (1993). Generating random binary deviates having fixed marginal distributions and specified degrees of association. *Amer. Statist.* 47:209-215.

[12] Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73:13-22.

[13] Liu, I. and Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics* 52:1223-1234.

[14] Liu, I. (2003). Describing ordinal odds ratios for stratified $r \times c$ tables. *Biometrical J.* 45:730-750.

[15] R Development Core Team (2007). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. ISBN 3-900051-07-0. http://www.R-project.org.

[16] Yan, J. and Fine, J. P. (2004). Estimating Equations for Association Structures *Statistics in Medicine* 23: 859-880.

[17] Yan, J. (2002). geepack: Yet Another Package for Generalized Estimating Equations *R-News* 2/3: 12-14.

APPENDIX

For convenience, we let $n'_{ik} := n_{ik} - 1$ and we often suppress subscripts i and k.

We can express $\pi^{00}_{xy|ik}$, $\pi^{01}_{xy|ik}$ and $\pi^{10}_{xy|ik}$ in terms of $\pi_{x|ik}$, $\pi_{y|ik}$ and $\pi_{xy|ik} := \pi^{11}_{xy|ik}$:

$$\begin{aligned}
\pi^{10}_{xy|ik} &= \pi_{x|ik} - \pi^{11}_{xy|ik} \\
\pi^{01}_{xy|ik} &= \pi_{y|ik} - \pi^{11}_{xy|ik} \\
\pi^{00}_{xy|ik} &= 1 - \pi_{x|ik} - \pi_{y|ik} + \pi^{11}_{xy|ik}
\end{aligned}$$

We can compute

$$\begin{aligned}
\mathrm{E}X_x X_y &= nn'\pi_x\pi_y + n\pi^{11}_{xy} \\
\mathrm{E}X_x \bar{X}_y &= nn'\pi_x\bar{\pi}_y + n\pi^{01}_{xy} \\
\mathrm{E}\bar{X}_x X_y &= nn'\bar{\pi}_x\pi_y + n\pi^{10}_{xy} \\
\mathrm{E}\bar{X}_x \bar{X}_y &= nn'\bar{\pi}_x\bar{\pi}_y + n\pi^{00}_{xy}
\end{aligned} \tag{4}$$

Theorem:
$U_{xy|ab}$ and $U_{xy|abc}$ are dually consistent.

Proof The proof is similar to Greenland (1989). Under either limiting model, we have

$$\begin{aligned}
\mathrm{Cov}(L^x_{ab}, L^y_{ab}) &= 1/(\Psi^x_{ab}\Psi^y_{ab})\mathrm{Cov}(\hat{\Psi}^x_{ab}, \hat{\Psi}^y_{ab}) \\
&= 1/(\Psi^x_{ab}\Psi^y_{ab})\frac{\mathrm{Cov}(C_{x|ab}-\Psi^x_{ab}C_{x|ba}, C_{y|ab}-\Psi^y_{ab}C_{y|ba})}{\mathrm{E}C_{x|ba}\mathrm{E}C_{y|ba}} \\
&= \frac{\mathrm{Cov}(C_{x|ab}-\Psi^x_{ab}C_{x|ba}, C_{y|ab}-\Psi^y_{ab}C_{y|ba})}{\mathrm{E}C_{x|ab}\mathrm{E}C_{y|ab}} \\
&= \frac{\sum_k \mathrm{Cov}(c_{x|ab}-\Psi^x_{ab}c_{x|ba}, c_{y|ab}-\Psi^y_{ab}c_{y|ba})}{\mathrm{E}C_{x|ab}\mathrm{E}C_{y|ab}} \\
&= \frac{\sum_k \mathrm{E}(c_{x|ab}-\Psi^x_{ab}c_{x|ba})(c_{y|ab}-\Psi^y_{ab}c_{y|ba})}{\mathrm{E}C_{x|ab}\mathrm{E}C_{y|ab}} \\
&= \frac{\{D_{ab}-\Psi^x_{ab}D^x_{ab}-\Psi^y_{ab}D^y_{ab}+\Psi^x_{ab}\Psi^y_{ab}D^{xy}_{ab}\}}{\mathrm{E}C_{x|ab}\mathrm{E}C_{y|ab}} \\
&= \frac{D_{ab}}{\mathrm{E}C_{x|ab}\mathrm{E}C_{y|ab}} - \frac{D^x_{ab}}{\mathrm{E}C_{x|ba}\mathrm{E}C_{y|ab}} - \frac{D^y_{ab}}{\mathrm{E}C_{x|ab}\mathrm{E}C_{y|ba}} + \frac{D^{xy}_{ab}}{\mathrm{E}C_{x|ba}\mathrm{E}C_{y|ba}}
\end{aligned}$$

with $D = \sum_k d$

$$\begin{aligned}
d_{ab} &= \frac{n_a n_b}{N_k^2}\{n'_a \pi_{x|a}\pi_{y|a}\pi^{00}_{xy|b} + n'_b \pi^{11}_{xy|a}\bar{\pi}_{x|b}\bar{\pi}_{y|b} + \pi^{11}_{xy|a}\pi^{00}_{xy|b}\} \\
d^x_{ab} &= \frac{n_a n_b}{N_k^2}\{n'_a \bar{\pi}_{x|a}\pi_{y|a}\pi^{10}_{xy|b} + n'_b \pi^{01}_{xy|a}\pi_{x|b}\bar{\pi}_{y|b} + \pi^{01}_{xy|a}\pi^{10}_{xy|b}\} \\
d^y_{ab} &= \frac{n_a n_b}{N_k^2}\{n'_a \pi_{x|a}\bar{\pi}_{y|a}\pi^{01}_{xy|b} + n'_b \pi^{10}_{xy|a}\bar{\pi}_{x|b}\pi_{y|b} + \pi^{10}_{xy|a}\pi^{01}_{xy|b}\} \\
d^{xy}_{ab} &= \frac{n_a n_b}{N_k^2}\{n'_a \bar{\pi}_{x|a}\bar{\pi}_{y|a}\pi^{11}_{xy|b} + n'_b \pi^{10}_{xy|a}\bar{\pi}_{x|b}\pi_{y|b} + \pi^{00}_{xy|a}\pi^{11}_{xy|b}\}
\end{aligned}$$

The first equality follows from the delta method, the second by writing $\hat{\Psi}_{ab} - \Psi_{ab} = \frac{C_{ab} - \Psi_{ab} C_{ba}}{C_{ba}}$, the third and last by $\Psi^x_{ab} = \mathrm{E}c_{x|ab}/\mathrm{E}c_{x|ba} = 1/\Psi^x_{ba}$, the fourth by independence of the strata and the fifth by $\mathrm{E}(c_{x|ab} - \Psi^x_{ab} c_{x|ba}) = 0$. The sixth equality is shown by applying straightforward calculation of the expectations using (4).

Similarly

$$\begin{aligned}
\mathrm{Cov}(L^x_{ab}, L^y_{ac}) &= 1/(\Psi^x_{ab}\Psi^y_{ac})\mathrm{Cov}(\hat{\Psi}^x_{ab}, \hat{\Psi}^y_{ac}) \\
&= 1/(\Psi^x_{ab}\Psi^y_{ac})\frac{\mathrm{Cov}(C_{x|ab} - \Psi^x_{ab}C_{x|ba}, C_{y|ac} - \Psi^y_{ac}C_{y|ca})}{\mathrm{E}C_{x|ba}\mathrm{E}C_{y|ca}} \\
&= \frac{\mathrm{Cov}(C_{x|ab} - \Psi^x_{ab}C_{x|ba}, C_{y|ac} - \Psi^y_{ac}C_{y|ca})}{\mathrm{E}C_{x|ab}\mathrm{E}C_{y|ac}} \\
&= \frac{\sum_k \mathrm{Cov}(c_{x|ab} - \Psi^x_{ab}c_{x|ba}, c_{y|ac} - \Psi^y_{ac}c_{y|ca})}{\mathrm{E}C_{x|ab}\mathrm{E}C_{y|ac}} \\
&= \frac{\{D_{abc} - \Psi^x_{ab}D^x_{abc} - \Psi^y_{ac}D^y_{abc} + \Psi^x_{ab}\Psi^y_{ac}D^{xy}_{abc}\}}{\mathrm{E}C_{x|ab}\mathrm{E}C_{y|ac}} \\
&= \frac{D_{abc}}{\mathrm{E}C_{x|ab}\mathrm{E}C_{y|ac}} - \frac{D^x_{abc}}{\mathrm{E}C_{x|ba}\mathrm{E}C_{y|ac}} - \frac{D^y_{abc}}{\mathrm{E}C_{x|ab}\mathrm{E}C_{y|ca}} + \frac{D^{xy}_{abc}}{\mathrm{E}C_{x|ba}\mathrm{E}C_{y|ca}}
\end{aligned}$$

with

$$\begin{aligned}
d_{abc} &= \frac{n_a n_b n_c}{N_k^2}\pi^{11}_{xy|a}\bar{\pi}_{x|b}\bar{\pi}_{y|c} \\
d^x_{abc} &= \frac{n_a n_b n_c}{N_k^2}\pi^{01}_{xy|a}\pi_{x|b}\bar{\pi}_{y|c} \\
d^y_{abc} &= \frac{n_a n_b n_c}{N_k^2}\pi^{10}_{xy|a}\bar{\pi}_{x|b}\pi_{y|c} \\
d^{xy}_{abc} &= \frac{n_a n_b n_c}{N_k^2}\pi^{00}_{xy|a}\pi_{x|b}\pi_{y|c}
\end{aligned}$$

The dually consistency follows from the fact that the $\hat{D}$'s and $\hat{D}$'s converge to the same expressions for both limiting models and that the $C_{x|ab}$ are exactly unbiased estimators of their expectations, also under both limiting models.