ASSESSING AGREEMENT AMONG RATERS AND IDENTIFYING ATYPICAL RATERS USING A LOG-LINEAR MODELING APPROACH

by

Kari B. Kastango

BS, University of Massachusetts, Amherst, 1989

MS, University of Massachusetts, Amherst, 1992

Submitted to the Graduate Faculty of

Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2006

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This dissertation was presented

by

Kari B. Kastango

It was defended on

March 23, 2006

and approved by

Dissertation Advisor: Roslyn A. Stone, Ph.D.
Associate Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member: Sati Mazumdar, Ph.D.
Professor
Department of Biostatistics,
Graduate School of Public Health
University of Pittsburgh

Committee Member: Howard E. Rockette, Ph.D.
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member: Mary Amanda Dew, Ph.D.
Professor
Department of Psychiatry, Psychology, and Epidemiology
Graduate School of Public Health
University of Pittsburgh

Committee Member: Benoit H. Mulsant, MD
Professor
Department of Psychiatry
School of Medicine
University of Pittsburgh

Roslyn A Stone, Ph.D.

ASSESSING AGREEMENT AMONG RATERS AND IDENTIFYING ATYPICAL RATERS USING A LOG-LINEAR MODELING APPROACH

Kari. B. Kastango, PhD

University of Pittsburgh, 2006

When an outcome is rated by several raters, ensuring consistency across raters increases the reliability of the measurement. Tanner and Young (1985) proposed a general class of log-linear models to assess agreement among K raters and a rating scale with *C* nominal categories. Their methodology can be used to assess pair-wise agreement among three or more raters. Rogel et al. (1996, 1998) extended this work by assessing various patterns of agreement among rater sub-groups of size K-1. These models can be used to test the assumption of rater exchangeability. Although parameters from these models can be used to identify atypical raters, no formal inferential procedures are available. I propose a formal inferential approach that can be used to test the assumption of rater exchangeability and to identify an atypical rater. The global and heterogeneous partial agreement model is fit to the data and pair-wise comparisons of the *K* partial agreement parameters are made, adjusting the p-values for the multiple comparisons made. The heterogeneous partial agreement parameter that is constantly involved in the pair-wise comparisons that are statistically significant is distinguished. The premise is that, if there is an atypical rater, at least one heterogeneous partial agreement parameter will differ from at least one of the remaining K-1 partial agreement parameters. The approach is illustrated using published data from an intestinal biopsy rating study with six raters (Rogel et al., 1998). Overall Type I error and the power of the inferential approach to correctly identify atypical raters are assessed via simulation with rater sub-groups of size 5. The Bonferroni, Sidak, and Holm’s Step-down procedures using the Bonferroni and Sidak adjustments are used to control the overall Type I error. Being able to correctly identify an atypical rater, if present, and improving the consistency of ratings directly, influence the reliability of the measurement and the power of the study for a given sample size. Consequently, more informative studies can be conducted of interventions (e.g., behavioral, medicinal) that may have a significant positive impact on the public’s health.

TABLE OF CONTENTS

LIST OF TABLES

DEDICATION

In loving memory of my father, Captain Edward Kastango, a person who was atypical in his own right and who remains a consistent source of strength and inspiration for me.

To Tucker Knox and Maximilian Alex Kinne

To Emma Foley

ACKNOWLEDGMENTS

I would like to give special thanks to Dr. Roslyn Stone, Chair of my committee, for mentoring me through the dissertation process with vision, unwavering support, patience and friendship. I would also like to thank my committee members Dr. Sati Mazumdar, Dr. Mary Amanda Dew, Dr. Benoit Mulsant and Dr. Howard Rockette for their unwavering support, guidance, and invaluable input through this challenging undertaking.

I am extremely grateful to have worked closely with Dr. Bruce Pollock, Dr. Judith Saxton, Dr. Robert Sweet and Dr. Jules Rosen. My doctoral training has been tremendously enriched by them; their expertise and compassion for their patients are inspiring.

Special thanks to Susie Grasky, Marcia Kurs-Lasky, and Bernie for their kindness, friendship and laughter. It has made this journey all the more enjoyable. I would be remiss if I didn't acknowledging the impact of the friendships of Mary Williams, Kris Ruppert and Peter.

I want to thank my immediate (Carlson, Kinne, Kastango, Young) and extended family (Bell, Hwang, Knapp, Folan, and Agarwal). Your love and support are immeasurable.

1. INTRODUCTION

Unreliable or imprecise measurement of the primary outcome, whether continuous or categorical, limits the power of a study. One of the fundamental issues surrounding the design and analysis of a study involving a primary outcome measured by a subjective nominal rating scale by multiple raters is the reduced reliability of the measurement due to rater differences in rating the response. These differences can occur between raters at a single time point (inter-rater) or within raters (intra-rater) across time. The larger the amount of variability due to inter-rater or intra-rater differences, the greater the reduction in a study's power. It is difficult to demonstrate the benefit of new treatments (e.g., behavioral interventions, medicine) with insufficiently powered studies.

I focus on 'agreement', defined as the reproducibility of a categorical outcome. Although agreement is defined for both ordinal and nominal outcomes, the focus of the present work is on nominal outcomes. For nominal outcomes, such as the absence or presence of lesions determined by categorizing morphological features of biopsy specimens, each rater should have sufficient experience with the histological characteristics associated with the lesion for correct classification. At best, each specimen is objectively categorized with each rater using the same classification criteria. At worst, the ratings are highly subjective.

Summary measures of overall agreement, such as Cohen's Kappa (Cohen, 1960) assume that the raters are interchangeable. This statistic does not focus explicitly on the contributions of individual raters or groups of raters to the overall summary measure and cannot be used to identify atypical raters. There are situations, however, where identifying atypical raters is of particular interest.

My work focuses on identifying atypical raters for nominal data. That is, are there any raters who are inconsistent in their characterizations of the outcome with respect to the other raters? If so, how can they be identified? The log-linear modeling approach of Rogel et al. (1996, 1998) allows the assumption of rater exchangeability to be tested in a setting where *K* raters rate the same patients. They define parameters that are used to assess various patterns of agreement but do not address explicitly the identification of individual raters. The focus of this work was to formalize inferential approaches to identify atypical rater(s) within the framework of these models.

In chapter 2 I review two approaches to assess inter-rater agreement for nominal categorical data, summary statistics and log-linear modeling. I focused on the log-linear modeling approaches used by Tanner and Young (1985, JASA) and Rogel et al. (1996, 1998) to model inter-rater agreement and quantify the magnitude of inter-rater agreement. I consider the applicability of formal statistical inference to identify an atypical rater, and review relevant multiple comparison procedures for identifying atypical raters.

The inferential approach and the simulation study conducted to assess the Type I error and power of the approach are described in Chapter 3. The analysis of published data from an inter-rater agreement study involving six raters using these methods and the results of the simulation study are presented in Chapter 4. Discussion of the results and conclusions are in Chapter 5.

2. LITERATURE REVIEW

I review two basic approaches to assessing rater agreement for nominal categorical data. One approach focuses on the use of the summary statistic Kappa (Cohen, 1960). The second approach focuses on modeling the structure of agreement in the data using log-linear models (Tanner and Young, 1985; Rogel et al., 1996, 1998).

2.1 QUANTIFYING AGREEMENT USING THE SUMMARY STATISTIC KAPPA

The predominant summary statistic for assessing agreement involving categorical data is the Kappa (Cohen 1960) statistic. The statistic originated as a chance-corrected coefficient of agreement for a fixed pair of raters (K=2) rating the same patients using a nominal rating scale with two outcomes (C=2). It has since been generalized to situations involving (i) two raters, multiple categories, (ii) multiple raters, two categories, and (iii) multiple raters, multiple categories. Although the statistic has been defined for both ordinal and nominal categories, Kappa for situations (i) – (iii) is described in the context of the nominal case.

2.1.1. Two Raters, Binary Outcome

The statistic for two raters and a binary outcome is described as follows: Suppose two raters independently identify N slides as having cancer cells absent (0) or present (1). Each slide is allocated into one of the 2^2 cells as shown in Table 1. Let $x_{i_1 i_2}$ represent the number of slides assigned to category i by the first rater and to category i by the second rater where index *i* takes values 0 or 1. Let x_{ii} represent the number of slides assigned to category i by both raters, x_{i_1+} represent the total number of slides assigned to category i by the first rater, and x_{+i_2} represent the number of slides assigned to category i by the second rater.

Table 1. General Layout of a 2x2 Contingency Table Denoting Agreement

	Rater 2		
Rater 1	**0**	**1**	**Total**
0	x_{00}	x_{01}	x_{0+}
1	x_{10}	x_{11}	x_{1+}
Total	x_{+0}	x_{+1}	$x_{++} = N$

Kappa is defined $\mathrm{K} = \frac{P_0 - P_e}{1 - P_e}$, (Fleiss, Cohen, Everitt, 1969) where P_o is the observed proportion of agreement and P_e is the expected proportion of agreement by chance,

$$P_0 = \sum_{i=0}^{1} x_{i_1 i_2} \text{ and } P_e = \sum_{i=0}^{1} \left(\frac{x_{i_1+}}{N} \right) \left(\frac{x_{+i_2}}{N} \right).$$

Landis and Koch (1977) assigned the following degree of agreement for varying values of K:

$\mathrm{K} < 0$	Poor	$0.41 \le \mathrm{K} < 0.6$	Moderate
$0 \le \mathrm{K} < 0.2$	Slight	$0.61 \le \mathrm{K} < 0.8$	Substantial
$0.21 \le \mathrm{K} < 0.4$	Fair	$0.81 \le \mathrm{K} \le 1.0$	Almost

perfect.

2.1.2. Two Raters, Multiple (*C*) Nominal Outcomes

An example of multiple nominal outcomes would be stages of sleep (e.g., Wakefulness, Stage 1, Stage 2, Stage 3, Stage 4, and Stage REM). Suppose two raters independently classify N epochs (defined time intervals, e.g., 30 seconds) of physiological data into one of C nominal sleep stage categories. Each epoch is allocated into one of the C^2 cells as shown in Table 2. Let $x_{i_1 i_2}$ represent the number of epochs assigned to category *i* by the first rater and to category *i* by the second rater where index *i* takes values 0 to C-1. Let x_{ii} represent the number of epochs assigned to category *i* by both raters, $x_{i_1 +}$ represent the total number of epochs assigned to category *i* by the first rater, and x_{+i_2} represent the number of epochs assigned to category *i* by the second rater.

Table 2. General Layout of a Two Dimensional C x C Contingency Table Denoting Agreement

		Rater 2				
		0	**1**	**…i2…**	***C*-1**	**Total**
Rater 1	**0**	x_{00}	x_{01}	X_{0i2}	$x_{C\text{-}1\ C\text{-}1}$	x_{0+}
	1	x_{10}	x_{11}	X_{1i2}	$x_{C\text{-}1\ C\text{-}1}$	x_{1+}
	: **i1** **:**	: x_{i0} :	: x_{i1} :	: x_{i1j2} :	: $x_{C\text{-}1\ C\text{-}1}$	: x_{i+} :
	***C*-1**	$x_{C\text{-}1\ 0}$	$x_{C\text{-}1\ 1}$	$x_{C\text{-}1\ i2}$	$x_{C\text{-}1\ C\text{-}1}$	$x_{\ C\text{-}1\ +}$
	Total	x_{+0}	x_{+1}	x_{+j}	$x_{+\ C\text{-}1}$	x_{++} = N

Kappa is defined $\mathrm{K} = \frac{P_o - P_e}{1 - P_e}$, (Fleiss, Cohen, Everitt, 1969) where Po is the observed proportion of agreement and Pe is the expected proportion of agreement by chance,

$$P_0 = \frac{1}{N} \sum_{i=0}^{C-1} x_{i_1 i_2} \quad \text{and} \quad P_e = \sum_{i=0}^{C-1} \left(\frac{x_{i_1 +}}{N} \right) \left(\frac{x_{+i_2}}{N} \right)$$

2.1.3. Multiple (K) Raters, Binary Outcome

Fleiss (1981) generalized the original kappa to the situation where there are more than 2 raters (K>2). The generalization is made with the assumption that "the raters responsible for rating one subject are not assumed to be the same as those responsible for rating another (p.225, Fleiss, 1981)". This kappa is given by

$$\mathrm{K} = 1 - \frac{\sum_{i=1}^{n} \frac{x_i\,(m_i - x_i)}{m_i}}{n(\bar{m} - 1)\,\overline{pq}}$$

where

$n =$ number of subjects rated;
$m_{\mathrm{i}} =$ number of raters rating subject i;
$x_{\mathrm{i}} =$ number of positive ratings on subject i;
$m_{\mathrm{i}} - x_{\mathrm{i}} =$ number of negative ratings on subject i;
$\bar{m} =$ mean number of ratings per subject $= \sum_{i=1}^{n} \frac{m_{\mathrm{i}}}{n}$;
$\bar{p} =$ overall proportion of positive ratings $= \frac{\sum_{i=1}^{n} x_i}{n\bar{m}}$;
$\bar{q} =$ 1-$\bar{p}$; overall proportion of negative ratings.

Note that no differentiation between raters is made with respect to each rater's contribution to either the summary statistic or the number of positive (or negative) ratings. Kappa is a summary statistic under the assumption that the raters are exchangeable. Tanner and Young (1985) point out that the Kappa statistics are not sensitive to differences between observed and expected patterns of agreement, and Kappa's value is a function of the marginal distribution of the raters.

2.2 MODELING AGREEMENT USING LOG-LINEAR MODELS

The second general approach to assessing rater agreement is to model the pattern of agreement in the data using log-linear models. Log-linear modeling can be viewed as regression for count data displayed in a multi-way contingency table. Using this approach, the agreement pattern in the data can be parameterized. Excellent resources about log-linear modeling include Bishop, Fienberg, and Holland (1975) and Agresti (2002). I review the log-linear modeling literature related to assessing agreement among raters.

2.2.1. Notation

Log-linear modeling can used to describe and make inferences about the patterns of association among the categorical variables in a multi-dimensional contingency table. The dimension of the contingency table depends upon the number of categorical variables of interest. In the agreement framework, where all K raters rate an outcome variable with C categories, the data can be displayed in a K-way C^K contingency table. The relevant literature has been developed in detail for the case of a binary outcome, although the approach generalizes directly for C>2 categories. I review the methods in the context of a binary outcome.

For example, the cross-classification of three raters (K=3) assigning a rating of '0' for the absence or '1' for the presence a symptom (C=2) can be presented in a three-way 2^3 contingency table as shown in Table 3. Extending the notation introduced in section 2.1.1 for a two-way C^2 contingency table, a third subscript (k) is needed to represent the cells of a three-way contingency table. Subscripts i_1, i_2, and i_3 represent the rating assigned by rater 1, rater 2, and rater 3, respectively and $(i_1\ i_2\ i_3)$ represent a rating pattern. Therefore, $x_{i_1 i_2 i_3}$ represents the number of patients assigned to category i_1 by the first rater, to category i_2 by the second rater, and to category i_3 by the third rater, where each of i_1, i_2, and i_3 is 0 or 1. For example, x_{010} represents the number of patients rated as a “0” by raters 1 and 3 and a “1” by rater 2. In addition, x_{i_1++}, x_{+i_2+}, and x_{++i_3} represent the marginal totals for each rater and $p_{0\cdot\cdot}, p_{\cdot0\cdot}$, and $p_{\cdot\cdot0}$ represent the marginal proportions of rater 1, rater 2 and rater 3, respectively, (i.e., the proportion of patients rater 1, rater 2, and rater 3 rated as a “0”).

Table 3. General Layout of a Three-Way 2^3 Contingency Table

		Rater 1 (*i_1*)			
		(*i_1*=0)		(*i_1*=1)	
		Rater 2 (*i_2*)		Rater 2 (*i_2*)	
		(*i_2=0*)	(*i_2=1*)	(*i_2=0*)	(*i_2=1*)
Rater 3	(*i_3=0*)	x_{000}	x_{010}	x_{100}	x_{110}
(*i_3*)	(*i_3=1*)	x_{001}	x_{011}	x_{101}	x_{111}

Count data from this inter-rater agreement study are assumed to follow a multinomial distribution, because a fixed number of patients (N) are classified according to the ratings of the K raters. I am interested in the joint distribution of the ratings. The probability that a rating pattern is $(i_1\ i_2\ i_3)$ is given by the density function

$$f(\{x_{i_1i_2i_3}\}) = \frac{N!}{\prod_{i_1,i_2,i_3} x_{i_1i_2i_3}!} \prod_{i_1,i_2,i_3} \left(\frac{m_{i_1i_2i_3}}{N}\right)^{x_{i_1i_2i_3}} \qquad \text{(eq. 2.1.1, Bishop et al., 1975)}$$

where $m_{i_1i_2i_3}$ represents the expected frequency of rating pattern $(i_1\ i_2\ i_3)$. The maximum likelihood estimate of the frequency of observed rating pattern $(i_1\ i_2\ i_3)$, $\hat{m}_{i_1i_2i_3}$, is a function of the minimal sufficient statistics, a set of marginal totals from the contingency table that depend on the hypothesized log-linear model (Bishop, 1975). For example, $x_{i++}, x_{+j+}, and\ x_{++k}$ are minimal sufficient statistics for $m_{i++}, m_{+j+},$ and m_{++k}, respectively, and $\hat{m}_{i_1i_2i_3} = \dfrac{(x_{i++})\times(x_{+j+})\times(x_{++k})}{N^2}$, under the model of independence.

2.2.2. Model Of Independence

For the case of three raters and a binary outcome, under the assumption of independence the expected number of agreements among the three raters is given by

$$m_{i_1i_2i_3} = \frac{x_{i_1++}}{N}\frac{x_{+i_2+}}{N}\frac{x_{++i_3}}{N}N \qquad \text{(eq. 2.1.2, Bishop et al., 1975)}$$

Taking the natural logarithm of this equation, the log-linear model of independence is given by

$$\log m_{i_1i_2i_3} = \log x_{i_1++} + \log x_{+i_2+} + \log x_{++i_3} - 2\log N \qquad \text{(eq. 2.1.3)}$$

Using the notation of Rogel et al. (1998) this equation can be rewritten as

$$\log m_{i_1i_2i_3} = m + l_{i_1}^{O_1} + l_{i_2}^{O_2} + l_{i_3}^{O_3}\ , \qquad \text{(eq. 2.1.4)}$$

where $m_{i_1i_2i_3}$ is the expected cell count (assumed to be strictly positive) in the $(i_1i_2i_3)^{th}$ cell, μ represents the overall effect, $l_{i_1}^{O_1}$ represents the effect due to the *i*th level of the first rater, $l_{i_2}^{O_2}$ represents the effect due to the *i*th level of the second rater, and $l_{i_3}^{O_3}$ represents the effect due to the *i*th level of the third rater. The *i*th level of a rater refers to the rating category assigned, here 0 or 1. The notation 'O$_p$' is used to denote the rater (observer) and *p* indexes the raters, *p* = 1 to K (here K=3). For *C possible* categories of the rating, the overall effect, μ , and each rater effect, $l_{i_p}^{O_p}$, are defined as follows:

$$m = \frac{1}{C*C*C}\sum_{i_1=0}^{C-1}\sum_{i_2=0}^{C-1}\sum_{i_3=0}^{C-1}\log m_{i_1i_2i_3}, \quad l_{\mathrm{i}_1}^{O_1} = \frac{1}{C*C}\sum_{i_2=0}^{C-1}\sum_{i_3=0}^{C-1}\log m_{i_1i_2i_3} - m, \quad l_{\mathrm{i}_2}^{O_2} = \frac{1}{C*C}\sum_{i_1=0}^{C-1}\sum_{i_3=0}^{C-1}\log m_{i_1i_2i_3} - m,$$

$$l_{\mathrm{i}_3}^{O_3} = \frac{1}{C*C}\sum_{i_1=0}^{C-1}\sum_{i_2=0}^{C-1}\log m_{i_1i_2i_3} - m, \quad \text{with } \sum_{i_1=0}^{C-1} l_{i_1}^{O_1} = \sum_{i_2=0}^{C-1} l_{i_2}^{O_2} = \sum_{i_3=0}^{C-1} l_{i_3}^{O_3} = 0. \qquad \text{(eq.2.1.5)}$$

The above ‘sum to zero’ constraints rather than ‘baseline’ constraints yields an interpretation of the lambda parameters with respect to average agreement rather than agreement in reference to a given rater. A rater effect at the i^{th} level is interpreted as the departure of the rater’s i^{th} category marginal mean from the overall mean. The model has (C-1)(C-1)(C-1) residual degrees of freedom. The model of independence (eq. 2.1.4) models agreement due to chance and allows for marginal homogeneity or marginal heterogeneity across raters. Marginal homogeneity means that the proportion of patients to which each rater assigns a given category is the same for all raters. Marginal heterogeneity means that the proportion of patients to which each rater assigns a given category is not the same for all raters. Marginal homogeneity or marginal heterogeneity can occur in models of independence as well in models of agreement.

2.2.3. Models Of Quasi-Independence

Experienced raters trained in the use of a rating scale would be expected to agree more often than not. That is, a greater number of counts would be expected along the main diagonal than an independence model would indicate. Quasi-independence describes a rating pattern configuration with no structure specified in the off-diagonal cells but a larger number of counts on the main diagonal than would be expected under independence.

Tanner and Young (JASA, 1985) laid the foundation for using log-linear models to assess rater agreement by proposing the use of the quasi-independence model, a general class of models of the form

$$\log m_{i_1 i_2 \ldots i_K} = m + l_{i_1}^{O_1} + l_{i_2}^{O_2} + \ldots + l_{i_K}^{O_K} + d_{i_1 i_2 \ldots i_K}\ , \qquad \text{(eq. 2.1.6)}$$

where $d_{i_1 i_2 \ldots i_K}$ can be composed of more than one value. This model was formally introduced into the statistical literature by Goodman (1968).

In the context of inter-rater agreement, the $d_{i_1 i_2 \ldots i_K}$ term represents rater agreement different than what would be expected by chance. The parameterization of $d_{i_1 i_2 \ldots i_K}$ specifies the raters considered and the pattern of agreement among those raters. For example, the $d_{i_1 i_2 \ldots i_K}$ term can denote which raters are considered in a rater subgroup or whether the level of agreement depends on the category of the outcome.

2.2.3.1. Agreement Among All Raters, Homogeneous Agreement Across Categories

The simplest log-linear model of quasi-independence is homogeneous agreement across categories as well as raters. For the case of three raters and a binary outcome, this model (Tanner and Young, 1985) is given by

$$\log m_{i_1 i_2 i_3} = m + l_{i_1}^{O_1} + l_{i_2}^{O_2} + l_{i_3}^{O_3} + d_{i_1 i_2 i_3}, \text{(eq.2.1.7)}$$

$$\text{where } d_{i_1 i_2 i_3} = [I_1 d],\ and\ I_1 = 1 \text{ if } i_1 = i_2 = i_3,\ \text{for } i = 0,1$$

$$= 0 \text{ otherwise}.$$

This parameterization of $d_{i_1 i_2 i_3}$ uses one parameter to denote agreement among the three raters and does not distinguish whether the agreement is on the absence or presence of the symptom.

The indicator variable I_1 equals one for rating patterns (000) and (111) and equals zero for any other possible rating pattern.

AGREEMENT AMONG ALL RATERS, HETEROGENEOUS AGREEMENT ACROSS CATEGORIES

One extension of equation (2.1.7) is to allow the level of agreement to differ by category of the rating. This model is called “heterogeneous agreement across categories”. For a binary rating, $d_{i_1 i_2 i_3}$ is defined using C=2 δ parameters to denote agreement among the three raters; separate parameters denote whether the agreement is on the absence or presence of the symptom. That is,

$$d_{i_1 i_2 i_3} = \left[I_0 d_0 \quad I_1 d_1 \right]', \quad \text{(eq.2.1.8)}$$

where $I_0 = 1$ if $i_1 = i_2 = i_3$ for $i = 0$, and 0 otherwise,

and $I_1 = 1$ if $i_1 = i_2 = i_3$, for $i = 1$, and 0 otherwise.

The indicator variables I_0 and I_1 equal one for rating patterns (000) and (111), respectively, and both equal zero for all other rating patterns. For the general case of C categories and K raters the $d_{i_1 i_2 \ldots i_K}$ term will contain C parameters.

AGREEMENT WITHIN SUBGROUPS OF RATERS

Tanner and Young (1985) investigated agreement among sub-groups of raters of size G (2 < G < K), called “G-tuples of raters” with $u = \binom{K}{G}$ distinct subgroups of size G. The $d_{i_1 i_2 \ldots i_K}$ term is defined by a set of parameters that represents the cells corresponding to agreement between a given subgroup of the K raters. When the size of the rater subgroup equals the total number of raters (i.e., $G = K$), the concern is whether the agreement is homogeneous or heterogeneous across categories of the rating. However, when the size of the rater subgroup is less the total number of raters, (i.e., $G < K$), two characteristics of agreement must be considered: (i) the pattern of agreement among the raters in each subgroup, and (ii) the pattern of agreement across categories. Since either can be homogeneous or heterogeneous, four scenarios are possible.

For example, in the case of three raters there is only one sub-group of size three (rater triplet, $G = 3, K = 3$) and agreement among the three raters is either homogeneous or heterogeneous across categories as described above in sections 2.2.3.1 and 2.2.3.2. When the size of the rater sub-group is two ($G = 2, K = 3$), there are three pairs of raters. Agreement among these three pairs of raters can be either; (i) homogeneous among rater pairs and homogeneous across categories, (ii) homogeneous among rater pairs and heterogeneous across categories, (iii) heterogeneous among rater pairs and homogeneous across categories, or (iv) heterogeneous among rater pairs and heterogeneous across categories. The parameterizations of $d_{i_1 i_2 i_3}$ for these four scenarios are given in the following four sub-sections.

Homogeneous Agreement Across Rater Subgroups, Homogeneous Across Categories

Homogeneous agreement across rater subgroups and homogeneous across categories is parameterized by using a single delta parameter to represent agreement within each subgroup. It denotes agreement among any of the υ subgroups of raters of size G for any category C of the rating scale. For the case of three raters and a binary outcome, homogeneous agreement across rater pairs and homogeneous across categories is defined as

$$d_{i_1 i_2 i_3} = [I_1 d_1] \qquad \text{(equation 2.1.9)}$$

where $I_1 = 1$ if $i_1 = i_2 \neq i_3,\ i = 0$ or 1, or if $i_1 = i_3 \neq i_2$, $i = 0$ or 1,

or if $i_2 = i_3 \neq i_1$, $i = 0$ or 1,

and 0 otherwise.

The indicator variable I_1 equals one for rating patterns (001), (110), (010), (101), (100) and (011), the rating patterns in which exactly one rater disagrees with the other raters.

<u>Homogeneous Agreement Across Rater Subgroups, Heterogeneous Across Categories</u>

Homogeneous agreement across rater subgroups and heterogeneous across categories is parameterized by using C delta parameters, one δ term to denote agreement among any of the υ subgroup of raters of size G for each category of the response. For the case of three raters and a binary outcome, homogeneous agreement across rater pairs and heterogeneous across categories is defined as

$$d_{i_1 i_2 i_3} = [I_1 d_1 \quad I_2 d_2] \qquad \text{(equation 2.1.10)}$$

where $I_1 = 1$ if $i_1 = i_2 = 0, i_3 = 1$,

or if $i_1 = i_3 = 0, i_2 = 1$,

or if $i_2 = i_3 = 0, i_1 = 1$,

and 0 otherwise,

and where $I_2 = 1$ if $i_1 = i_2 = 1, i_3 = 0$,

or if $i_1 = i_3 = 1, i_2 = 0$,

or if $i_2 = i_3 = 1, i_1 = 0$,

and 0 otherwise.

The parameter d_1 represents rating patterns (001), (010) and (100) whereas d_2 represents rating patterns (110), (101) and (011).

Heterogeneous Agreement Across Rater Subgroups, Homogeneous Across Categories

Heterogeneous agreement across rater subgroups and homogeneous across categories is parameterized by using υ delta parameters, with one δ term to denote agreement between each of the υ subgroup of raters of size G. The parameterization of δ does not depend upon the category of the rating. For the case of three raters and a binary outcome, heterogeneous agreement across rater pairs and homogeneous across categories is defined as

$$d_{i_1 i_2 i_3} = \left[I_1 d_1 \quad I_2 d_2 \quad I_3 d_3 \right] \qquad \text{(equation 2.1.11)}$$

where $I_1 = 1$ if $i_1 = i_2 \neq i_3, i = 0$ or 1, and 0 otherwise,

where $I_2 = 1$ if $i_1 = i_3 \neq i_2, i = 0$ or 1, and 0 otherwise,

where $I_3 = 1$ if $i_2 = i_3 \neq i_1, i = 0$ or 1, and 0 otherwise.

The indicator variable I_1 equals one for rating patterns (001) and (110) signifying agreement between raters 1 and 2, I_2 equals one for rating patterns (010) and (101) signifying agreement between raters 1 and 3, and I_3 equals one for rating patterns (100) and (011) signifying agreement between raters 2 and 3. Each delta term represents a distinct subgroup of cells. No cell of the contingency table is parameterized by more than one of these delta parameters, because separate parameters are defined to reflect agreement within subgroups larger than G.

Heterogeneous Agreement Across Rater Subgroups, Heterogeneous Across Categories

Heterogeneous agreement across rater subgroups and heterogeneous across categories is parameterized by using C delta parameters for each of the υ subgroups of raters of size G. The parameterization of δ does distinguish the level of the rating. For the case of three raters and a binary outcome, the parameterization of $\boldsymbol{d}_{i_1 i_2 i_3}$ for heterogeneous agreement across rater pairs and heterogeneous across categories is defined as

$$\boldsymbol{d}_{i_1 i_2 i_3} = \left[I_1 \boldsymbol{d}_1 \quad I_2 \boldsymbol{d}_2 \quad I_3 \boldsymbol{d}_3 \quad I_4 \boldsymbol{d}_4 \quad I_5 \boldsymbol{d}_5 \quad I_6 \boldsymbol{d}_6 \right] \quad \text{(equation 2.1.12)}$$

where $I_1 = 1$ if $i_1 = i_2 = 0, i_3 = 1$, and 0 otherwise, and

where $I_2 = 1$ if $i_1 = i_2 = 1, i_3 = 0$, and 0 otherwise, and

where $I_3 = 1$ if $i_1 = i_3 = 0, i_2 = 1$, and 0 otherwise, and

where $I_4 = 1$ if $i_1 = i_3 = 1, i_2 = 0$, and 0 otherwise, and

where $I_5 = 1$ if $i_2 = i_3 = 0, i_1 = 1$, and 0 otherwise, and

where $I_6 = 1$ if $i_2 = i_3 = 1, i_1 = 0$, and 0 otherwise.

The indicator variable I_1 equals one for rating pattern (001), agreement between raters 1 and 2 on the absence of the symptom and indicator variable I_2 equals one for rating pattern (110), agreement between raters 1 and 2 on the presence of the symptom. Indicator variables I_3 and I_4, and I_5 and I_6, are defined similarly to denote pair-wise agreement between raters 1 and 3, and raters 2 and 3, respectively. With this parameterization, each delta term represents a distinct cell. No cell of the contingency table is parameterized by more than one of these delta parameters.

Tanner and Young (1985) defined homogeneous and heterogeneous agreement across C categories among subgroups of size G for K raters, but their examples were limited to scenarios of (i) two-rater, three-category outcome and (ii) three-rater, binary outcome. They also considered comparisons to a gold standard. They did not focus on the use of such parameterizations to identify atypical raters.

2.3 GLOBAL AND PARTIAL AGREEMENT

Rogel et al. (1996, 1998) extended this log-linear model approach to assess agreement among subgroups of K raters to the problem of identifying atypical raters by modeling agreement among rater subgroups of decreasing size. Although they developed the approach for the general case of K raters, they illustrated the approach in the context of a binary outcome (with six raters) and subgroups of size K-1. They introduced a 'global' and 'partial' agreement terminology in the framework of quasi-independence models. Global agreement is defined as agreement among all K raters and partial agreement is agreement among sub-groups of raters of size s where 2 < s < K.

Rogel et al. (1996, 1998) introduced the notation S_s, $S_{s,i}$ and $d_{K-1}^{G_{\bar{p}}}$ to describe explicitly how the agreement parameters are defined. S_s denotes the set of rating patterns where exactly s raters agree regardless of category. For example, the set of rating patterns representing homogeneous agreement across categories among six raters is denoted by S_6, and S_5 denotes the set of rating patterns representing homogeneous agreement across categories among sub-groups of five raters. $S_{s,i}$ denotes heterogeneous agreement across categories ($i = 0$ to C) among rater sub-groups of size s. For example,

$S_{5,0}$ denotes the rating patterns where exactly five raters agree on category 0. Lastly, $d_{K-1}^{\overline{p}}$ identifies which rater is omitted from the rater sub-group where p is the rater index ($p = 1$ to K). For K=6 and a rater subgroup of size five, homogeneous agreement across categories with homogeneous agreement among all raters but rater 3 is denoted by $d_{i_1i_2i_3i_4i_5i_6} = d_5^{G_{\overline{3}}}$ and heterogeneous across categories with homogeneous agreement among all raters but rater 3 is denoted by $d_{i_1i_2i_3i_4i_5i_6} = d_{5,i}^{G_{\overline{3}}}$ ($i = 0,...,C-1$). Using this notation, I review five quasi-independent log-linear models used in their global and partial agreement modeling approach.

For the case of six raters and a binary outcome, the general form of the quasi-independence model is

$$\log m^{QI}_{i_1i_2i_3i_4i_5i_6} = m + l^{O_1}_{i_1} + l^{O_2}_{i_2} + l^{O_3}_{i_3} + l^{O_4}_{i_4} + l^{O_5}_{i_5} + l^{O_6}_{i_6} + d_{i_1i_2i_3i_4i_5i_6} \qquad \text{(equation 2.1.13)}$$

where μ represents the overall effect, $l^{O_p}_{i_p}$ represents the effect of observer O_p (p=1,2,…,6) on category i_p (i_p = 0,1) with 'sum to zero' constraints on the $l^{O_p}_{i_p}$ terms. In this model, $d_{i_1i_2i_3i_4i_5i_6}$ represents rater agreement different than what would be expected by chance.

2.3.1. Global Agreement, Homogeneous Across Categories (G)

The simplest log-linear model of quasi-independence for six raters and a binary outcome is homogeneous agreement across categories. This model is referred to as the global agreement model 'G' by (Rogel et al., 1996, 1998)

$$\log m^{G}_{i_1i_2i_3i_4i_5i_6} = m + l^{O_1}_{i_1} + l^{O_2}_{i_2} + l^{O_3}_{i_3} + l^{O_4}_{i_4} + l^{O_5}_{i_5} + l^{O_6}_{i_6} + d_{i_1i_2i_3i_4i_5i_6}$$

$$\text{and } d_{i_1 i_2 i_3 i_4 i_5 i_6} = \left[I_1 d_6 \right] \quad \text{(equation 2.1.14)}$$

$$\text{where } I_1 = 1 \text{ if } (i_1, i_2, \ldots, i_6) \in S_6 \text{ and 0 otherwise.}$$

S_6 denotes the set of rating patterns representing agreement among all six raters. The indicator variable $I_1 = 1$ denotes rating patterns (000000) and (111111). This is analogous to the agreement among all raters, homogeneous agreement across categories of Tanner and Young (1985).

2.3.2. Global Agreement, Heterogeneous Across Categories (Gc)

Equation (2.13) allows the level of agreement to differ by category. Rogel et al (1996) call this model 'Gc' and it is given by

$$\log m^{Gc}_{i_1 i_2 i_3 i_4 i_5 i_6} = m + l^{O_1}_{i_1} + l^{O_2}_{i_2} + l^{O_3}_{i_3} + l^{O_4}_{i_4} + l^{O_5}_{i_5} + l^{O_6}_{i_6} + d_{i_1 i_2 i_3 i_4 i_5 i_6}$$

$$\text{and } d_{i_1 i_2 i_3 i_4 i_5 i_6} = [\ \left[I_1 d_{6,0} \quad I_2 d_{6,1} \right] \qquad \text{(equation 2.1.15)}$$

$$\text{where } I_1 = 1 \text{ if } (i_1, i_2, \ldots, i_6) \in S_{6,0}\text{; 0 otherwise,}$$

$$\text{where } I_2 = 1 \text{ if } (i_1, i_2, \ldots, i_6) \in S_{6,1}\text{; 0 otherwise.}$$

The $d_{6,0}$ term denotes the rating pattern, (000000), where all six observers agree on category '0' and $d_{6,1}$ denotes the rating pattern, (111111), where all six observers agree on category '1'. This is analogous to the model of agreement among all raters, heterogeneous across categories, of Tanner and Young (1985).

Rogel et al. (1996, 1998) also introduced two 'global and partial agreement' models. These models parameterize the $d_{i_1 i_2 i_3 i_4 i_5 i_6}$ term to assess the agreement structure among all six (K) raters (global agreement) and among rater sub-groups of size five (K-1 partial agreement). Rogel et al. (1996, 1998) refer to the ‘global and partial agreement’ model as ‘GP’ if the global and partial agreements are homogeneous according to categories of the ratings and ‘GPc’ if the global and partial agreements are heterogeneous according to categories of the ratings.

2.3.3. Global And Partial Agreement, Homogenous Across Categories (GP)

The GP model describing homogeneous agreement across categories is given by

$$\log m^{GP}_{i_1i_2i_3i_4i_5i_6} = m + l^{O_1}_{i_1} + l^{O_2}_{i_2} + l^{O_3}_{i_3} + l^{O_4}_{i_4} + l^{O_5}_{i_5} + l^{O_6}_{i_6} + d_{i_1i_2i_3i_4i_5i_6}$$

$$\text{and } d_{i_1i_2i_3i_4i_5i_6} = [\ \left[I_1 d_6 \quad I_2 d_5 \right] \quad \text{(equation 2.1.16)}$$

$$\text{where} \quad I_1 = 1 \text{ if } (i_1, i_2, ..., i_6) \in S_6,$$

$$I_2 = 1 \text{ if } (i_1, i_2, ..., i_6) \in S_5,$$

and 0 otherwise.

The d_6 term represents rating patterns {(000000), (111111)} and the d_5 term represents rating patterns {(000001), (111110), (000010), (111101), (000100), (111011), (001000),(110111), (010000), (101111), (100000), and (011111)}, patterns with exactly one disagreement.

2.3.4. Global And Partial Agreement, Heterogeneous Across Categories (GPc)

The GPc model describing heterogeneous agreement across categories is given by

$$\log m^{GPc}_{i_1i_2i_3i_4i_5i_6} = m + l^{O_1}_{i_1} + l^{O_2}_{i_2} + l^{O_3}_{i_3} + l^{O_4}_{i_4} + l^{O_5}_{i_5} + l^{O_6}_{i_6} + d_{i_1i_2i_3i_4i_5i_6}$$

$$\text{and} \quad d_{i_1i_2i_3i_4i_5i_6} = \left[I_1 d_{6,0} \quad I_2 d_{6,1} \quad I_3 d_{5,0} \quad I_4 d_{5,1} \right] \quad \text{(equation 2.1.17)}$$

$$\text{where} \quad I_1 = 1 \text{ if } (i_1, i_2, ..., i_6) \in S_{6,0},$$

$$I_2 = 1 \text{ if } (i_1, i_2, ..., i_6) \in S_{6,1},$$

$$I_3 = 1 \text{ if } (i_1, i_2, ..., i_6) \in S_{5,0},$$

$$I_4 = 1 \text{ if } (i_1, i_2, ..., i_6) \in S_{5,1},$$

$$\text{and} \quad 0 \text{ otherwise.}$$

That is, $d_{6,0}$ denotes rating pattern (000000), $d_{6,1}$ denotes rating pattern (111111), $d_{5,0}$ denotes rating patterns {(000001), (000010), (000100), (001000), (010000), and (100000)} and $d_{5,1}$ denotes rating patterns {(111110), (111101), (111011), (110111), (101111), (011111)}. The terms $d_{5,0}$ and $d_{5,1}$ represent rating patterns where there is exactly one disagreement from the rating of 0 and 1, respectively.

2.3.5. Global And Heterogeneous Partial Agreement, Homogeneous Across Categories (GHeP)

Rogel et al. (1996, 1998) suggested that the global and partial agreement models could be used to identify atypical raters if one kept track of which rater was omitted in each subgroup of five raters. They presented a 'global and heterogeneous partial agreement' (GHeP) model that denotes differing levels of agreement among rater subgroups of size five that is homogeneous across categories of the ratings. The 'GHeP' model is defined as follows:

$$\log m^{GHeP}_{i_1i_2i_3i_4i_5i_6} = m + l^{O_1}_{i_1} + l^{O_2}_{i_2} + l^{O_3}_{i_3} + l^{O_4}_{i_4} + l^{O_5}_{i_5} + l^{O_6}_{i_6} + d_{i_1i_2i_3i_4i_5i_6}$$

and $$d_{i_1 i_2 i_3 i_4 i_5 i_6} = \left[\mathrm{I}_0 d_6 \quad \mathrm{I}_1 d_5^{G_{\bar{1}}} \quad \mathrm{I}_2 d_5^{G_{\bar{2}}} \quad \mathrm{I}_3 d_5^{G_{\bar{3}}} \quad \mathrm{I}_4 d_5^{G_{\bar{4}}} \quad \mathrm{I}_5 d_5^{G_{\bar{5}}} \quad \mathrm{I}_6 d_5^{G_{\bar{6}}} \right]' \text{ (equation 2.1.18)}$$

where $\mathrm{I}_0 = 1$ if $(i_1, i_2, \ldots, i_6) \in S_6$

$\mathrm{I}_1 = 1$ if $(i_2, \ldots, i_6) \in S_5$

$\mathrm{I}_2 = 1$ if $(i_1, i_3, \ldots, i_6) \in S_5$

$\mathrm{I}_3 = 1$ if $(i_1, i_2, i_4, i_5, i_6) \in S_5$

$\mathrm{I}_4 =1$ if $(i_1, i_2, i_3, i_5, i_6) \in S_5$

$\mathrm{I}_5 =1$ if $(i_1, i_2, i_3, i_4, i_6) \in S_5$

$\mathrm{I}_6 =1$ if $(i_1, \ldots, i_5) \in S_5$

and 0 otherwise.

The term $d_5^{G_{\bar{p}}}$ denotes the level of agreement in the rater sub-group of exactly five observers after removal rater *p* (*p*= 1,…,6). For example, $d_5^{G_{\bar{3}}}$ represents rating patterns {(001000), (110111)}, i.e., homogeneous agreement across categories among all raters except Rater 3.

2.4 AN EXAMPLE OF MODELING AGREEMENT USING LOG-LINEAR MODELS: INTESTINAL BIOPSY RATING DATA

The use of log-linear models to model agreement is illustrated using the published data of Rogel et al. (1998), in which six raters assessed the absence (rating of 0) or presence (rating of 1) of mucosecretion diminution.

The 25 rating patterns observed, the frequency of each pattern and the type of agreement represented by each rating pattern are shown in Table 4. For the 68 biopsies, exact six-way agreement was observed for 30 and exact five-way agreement for 17, with

21 biopsies having other rating patterns. Overall, 31.1% of ratings were for the presence of mucosecretion diminution. Rater 4 was in disagreement for 35% (6 of 17) of biopsies characterized by five-way agreement, while rater 2 and rater 6 each were in disagreement for 17% (3 of 17) of the biopsies characterized as having five-way agreement. The remaining 39 possible rating patterns with counts of zero are not listed in Table 4.

Table 4. Frequencies and Type of Agreement of the 25 Observed Rating Patterns from Six Raters Denoting the Absence (0) or Presence (1) of Mucosecretion Diminution in 68 Intestinal Biopsy Specimens

Rating Pattern*	**Count**	**Agreement Parameters ****
111111	1	δ_6 , $\delta_{6,1}$
111110	2	δ_5, $\delta_{5,1}$, $d_5^{G_{\bar{6}}}$
111101	2	δ_5, $\delta_{5,1}$, $d_5^{G_{\bar{5}}}$
111100	2	
110111	1	δ_5, $\delta_{5,1}$, $d_5^{G_{\bar{3}}}$
110101	1	
110100	1	
101111	2	δ_5, $\delta_{5,1}$, $d_5^{G_{\bar{2}}}$
101100	1	
100111	1	
100110	1	
011111	2	δ_5, $\delta_{5,1}$, $d_5^{G_{\bar{1}}}$
011110	2	
011101	1	
010111	3	
010101	1	
010100	1	
010000	1	δ_5, $\delta_{5,0}$, $d_5^{G_{\bar{2}}}$
001110	2	
001100	2	
000110	1	
000101	1	

Table 4 (continued)

Rating Pattern*	Count	Agreement Parameters **
000100	6	$\delta_5, \delta_{5,0}, d_5^{G_{\overline{4}}}$
000001	1	$\delta_5, \delta_{5,0}, d_5^{G_{\overline{6}}}$
000000	29	$\delta_6,\ \delta_{6,0}$

* Rating patterns not listed have frequencies of zero.

**Rating patterns without any indication of 'type of agreement' represent patterns having less than five raters in agreement on either the absence or presence of mucosecretion diminution.

These intestinal biopsy data are summarized in Table 5 in terms of the marginal percentages of the absence and presence of mucosecretion diminution by each rater and the percentage of ratings that exhibit each type of agreement. Rater 4 had the highest marginal percentage for the presence of mucosecretion diminution (54.4%). Global agreement was observed for 44.1% of the biopsies (including 42.6% on the absence and 1.5% on the presence of mucosecretion diminution). Five raters agreed on 25% of the biopsies (11.7% on absence and 13.2% on presence of mucosecretion diminution). The percentage of ratings showing partial agreement when each rater is excluded in turn is shown in the last column of Table 5. For example, the percentage of biopsies showing agreement when rater 4 is excluded is 6/68 = 8.8%. A higher level of partial agreement when a rater is excluded indicates that the excluded rater is in disagreement relatively more often when the other raters agree.

Table 5. Marginal Percentages for Different Category Specific Agreement Patterns and Rater Exclusion

Rater i	Marginal % for Absence	Marginal % for Presence	G %, d_6	G on Absence %, $d_{6,0}$	G on Presence %, $d_{6,1}$	GP %, d_5	GP on Absence %, $d_{5,0}$	GP on Absence %, $d_{5,1}$	Excluded Rater, % $d_5^{\bar{i}}$
1	77.9	22.1	44.1	42.6	1.5	25.0	11.7	13.2	2.9
2	69.1	30.9							4.4
3	72.1	27.9							1.5
4	45.6	54.4							8.8
5	73.5	26.5							2.9
6	75.0	25.0							4.4

* For agreement patterns see Table 4. G = Global agreement; GP= Global and partial agreement

For the intestinal biopsy data, the estimates of the rater effects, global agreement and the six partial agreement parameters are given in Table 6. Parameter estimates are given under the assumption of marginal homogeneity and marginal heterogeneity. Parameter estimates in Table 6 were derived using sum-to-zero constraints for the ? parameters (Appendix A) and indicator variables for the global and heterogeneous partial agreement parameters (Appendix B). Stata, version 8.2, was used to fit the models (Appendix C).

Table 6. Maximum Likelihood Estimates and Standard Errors of the Global and Heterogeneous Partial Agreement Parameters Assuming Marginal Homogeneity and Heterogeneity for the Intestinal Biopsy Data.

	Assumed Marginals				
	Homogeneous			**Heterogeneous**	
Parameter	Estimate (SE)	Wald Test p-value		Estimate (SE)	Wald Test p-value
l^{O_1}	na	na		-0.65 (0.20)	<0.01
l^{O_2}	na	na		-0.25 (0.17)	0.16
l^{O_3}	na	na		-0.35 (0.19)	0.07
l^{O_4}	na	na		1.35 (0.37)	<0.001
l^{O_5}	na	na		-0.42 (0.19)	0.02
l^{O_6}	na	na		-0.48 (0.18)	<0.01
d_6	3.58 (0.28)	<0.001		4.50 (0.54)	<0.001
$d_5^{\overline{1}}$	0.87 (0.74)	0.24		1.96 (0.85)	0.02
$d_5^{\overline{2}}$	1.27 (0.62)	0.04		2.44 (0.79)	<0.01
$d_5^{\overline{3}}$	0.17 (1.02)	0.87		1.38 (1.13)	0.22
$d_5^{\overline{4}}$	1.96 (0.46)	<0.001		0.36 (0.56)	0.52
$d_5^{\overline{5}}$	0.87 (0.74)	0.24		2.08 (0.87)	0.02
$d_5^{\overline{6}}$	1.27 (0.62)	0.04		2.47 (0.76)	<0.01
μ	-0.87 (0.21)	<0.001		-2.08 (0.51)	<0.001

na = not applicable

What does it mean to be atypical under the assumption of marginal homogeneity? Under the assumption of marginal homogeneity, the magnitude of each heterogeneous partial agreement parameter, $d_5^{\bar{i}}$, corresponds to each rater's non-chance contribution to five-way agreement after accounting for global agreement. The rater with the largest heterogeneous partial agreement parameter estimate is the rater who disagrees more often than the other five raters. The partial agreement parameters essentially partition the five-way agreement into components attributable to each rater.

In Table 6 for the assumption of marginal homogeneity, the heterogeneous partial agreement parameter estimate of 1.96 for Rater 4 reflects the six biopsy specimens where Rater 4 disagreed with the other five raters. Five-way agreement with Rater 4 being the discrepant rater is represented by rating patterns (000100) and (111011). The log-linear model for the expected number of counts is given by

$\log m_{000100} = -0.87 + 3.58 d_6 + 0.87 d_5^{\bar{1}} + 1.27 d_5^{\bar{2}} + 0.17 d_5^{\bar{3}} + 1.96 d_5^{\bar{4}} + 0.87 d_5^{\bar{5}} + 1.27 d_5^{\bar{6}}$ where $d_6, d_5^{\bar{1}}$, $d_5^{\bar{2}}, d_5^{\bar{3}}, d_5^{\bar{5}}$ and $d_5^{\bar{6}} = 0$, and $d_5^{\bar{4}} = 1$.

$\log m_{000100} = -0.87 + 1.96 = 1.01$

$m_{000100} = \exp(1.01) = 2.75$

The expected number of biopsy specimens with rating pattern (000100) is 2.75. Similarly, the expected number of biopsy specimens with rating pattern (111011) is 2.75. Therefore, under this model the total number of expected biopsy specimens having a rating pattern representing five-way agreement where Rater 4 is the disagreeing rater is 5.5 (compared to the 6 shown in Table 5). From the observed data, all six disagreements came from Rater 4 rating the presence of mucosecretion diminution when the remaining five raters rated the absence of mucosecretion diminution.

The heterogeneous partial agreement parameter estimate of 0.17 for Rater 1 reflects the one disagreement Rater 1 had with the remaining five raters. The expected number of counts is $2e^{(-0.87+0.17)} = 2e^{-0.7} = 2*0.496 = 0.99$. The standard error of the heterogeneous partial agreement parameter $d_5^{\bar{3}}$ is much larger than for any other parameter in the model, because Rater 3 was in disagreement only once with the other five raters whereas the remaining five raters where in disagreement three or more times when five-way agreement was considered. The heterogeneous partial agreement parameter estimates of 1.27 for Rater 2 and Rater 6 reflect the three disagreements each rater has with the remaining five raters.

In Table 6 for the assumption of marginal heterogeneity, Rater 4's lambda parameter estimate, l^{O_4} = 1.35, and heterogeneous partial agreement parameter estimate of 0.36 reflects the five times Rater 4 rated "presence" when the remaining five raters rated "absence" and that Rater 4 never rated "absence" when the remaining five raters rated "presence". The relatively large magnitude of l^{O_4} compared to the other five raters' l^{O_i} indicates that Rater's 4 marginal proportion for "presence", 54.4%, is higher than the overall mean portion for "presence", 31.3%. The log-linear model for the expected number of counts for rating pattern (000100) is given by

$$\log m_{000100} = -2.08 + l^{O_1}(-0.65) + l^{O_2}(-0.25) + l^{O_3}(-0.25) + l^{O_4}(1.35) + l^{O_5}(-0.42) +$$
$$l^{O_6}(-0.48) + d_6(4.50) + 1.96 d_5^{\bar{1}} + 2.44 d_5^{\bar{2}} + 1.38 d_5^{\bar{3}} + 0.36 d_5^{\bar{4}} + 2.08 d_5^{\bar{5}} + 2.47 d_5^{\bar{6}} \text{ where}$$
$$l^{O_1}, l^{O_2}, l^{O_3}, l^{O_5}, l^{O_6} = -1, \text{ and } l^{O_4} = 1, \text{ and } d_6, d_5^{\bar{1}}, d_5^{\bar{2}}, d_5^{\bar{3}}, d_5^{\bar{5}} \text{ and } d_5^{\bar{6}} = 0, \text{ and } d_5^{\bar{4}} = 1.$$
$$\log m_{000100} = -2.08 + 0.65 + 0.25 + 0.25 + 1.35 + 0.42 + 0.48 + 0.36 = 1.68.$$

The expected number of counts for rating pattern (000100), m_{000100}, is $\exp(1.68) = 3.75$.

The log-linear model for the expected number of counts for rating pattern (111011) is given by the above equation, but now $l^{O_4} = -1$, and $l^{O_1}, l^{O_2}, l^{O_3}, l^{O_5}, l^{O_6} = 1$.

Indicator variables $d_6, d_5^{\bar{1}}, d_5^{\bar{2}}, d_5^{\bar{3}}, d_5^{\bar{5}}$ and $d_5^{\bar{6}}$ still equal 0, and $d_5^{\bar{4}}$ equals 1.

$$\log m_{000100} = -2.08 - 0.65 - 0.25 - 0.25 - 1.35 - 0.42 - 0.48 + 0.36 = -5.12.$$

The expected number of counts for rating pattern (000100), m_{111011}, is $\exp(-5.12) = 0.005 = \sim 0$.

Table 7 summarizes the predicted counts for the G agreement, Gc agreement, GP agreement, GPc agreement and GHeP agreement models considered under the assumption of marginal homogeneity fitted to the intestinal biopsy data. For example, under the G model, 15 biopsies are predicted to be rated as having the absence of mucosecretion diminution and 15 biopsies are predicted to be rated as having the presence of mucosecretion diminution. Given that the overall expected number of ratings has to equal the observed number of ratings, the remaining 38 ratings are equally dispersed across the remaining 62 possible rating patterns, giving a predicted count of 0.61. The GPc model, the best fitting model under the assumption of marginal homogeneity, yields predicted cell counts that are closer to the observed cell counts for each of the 64 rating patterns. Note that 21 biopsies (30.8%) had one of the14 rating patterns that did not represent global or partial agreement.

Table 7. Predicted Counts of Observed Rating Pattern Based on Five Models under the Assumption of Marginal Homogeneity and Marginal Heterogeneity

	Marginal Homogeneity					
Rating Pattern	Observed Count	G	Gc	GP	GPc	GHeP
111111	1	15.00	1.00	15.00	1.00	15.00
111110	2	0.60	0.61	1.42	1.50	1.50
111101	2	0.63	0.61	1.42	1.50	1.00
111100	2	0.61	0.61	0.42	0.42	0.42
110111	1	0.61	0.61	1.42	1.50	0.5
110101	1	0.61	0.61	0.42	0.42	0.42
110100	1	0.61	0.61	0.42	0.42	0.42
101111	2	0.61	0.61	1.42	1.50	1.50
101100	1	0.61	0.61	0.42	0.42	0.42
100111	1	0.61	0.61	0.42	0.42	0.42
100110	1	0.61	0.61	0.42	0.42	0.42
011111	2	0.61	0.61	1.42	1.50	1.00

Table 7 (continued)

	Marginal Homogeneity					
Rating Pattern	Observed Count	G	Gc	GP	GPc	GHeP
011110	2	0.61	0.61	0.42	0.42	0.42
011101	1	0.61	0.61	0.42	0.42	0.42
010111	3	0.61	0.61	0.42	0.42	0.42
010101	1	0.61	0.61	0.42	0.42	0.42
010100	1	0.61	0.61	0.42	0.42	0.42
010000	1	0.61	0.61	1.42	1.33	1.50
001110	2	0.61	0.61	0.42	0.42	0.42
001100	2	0.61	0.61	0.42	0.42	0.42
000110	1	0.61	0.61	0.42	0.42	0.42
000101	1	0.61	0.61	0.42	0.42	0.42
000100	6	0.61	0.61	1.42	1.33	3.00
000001	1	0.61	0.61	1.42	1.33	1.50
000000	29	15.00	29.00	15.00	29.00	15.00
	Marginal Heterogeneity					
Rating Pattern	Observed Count	G	Gc	GP	GPc	GHeP
111111	1	4.76	1.00	5.92	1.00	5.03
111110	2	0.32	0.92	0.83	1.97	1.74
111101	2	0.28	0.83	0.72	1.74	1.05
111100	2	0.60	1.10	0.43	0.94	0.34
110111	1	0.02	0.04	0.07	0.06	0.01
110101	1	0.47	0.89	0.33	0.73	0.26
110100	1	1.02	1.19	0.69	0.76	0.69
101111	2	0.20	0.61	0.48	1.17	1.05
101100	1	0.82	0.96	0.53	0.59	0.56
100111	1	0.34	0.65	0.22	0.49	0.18
100110	1	0.73	0.87	0.46	0.51	0.48
011111	2	0.40	1.15	1.12	2.52	1.46
011110	2	0.86	1.52	0.68	1.37	0.53
011101	1	0.77	1.37	0.59	1.21	0.47
010111	3	0.69	1.23	0.51	1.07	0.41
010101	1	1.31	1.47	0.94	0.98	0.96
010100	1	2.8	1.95	2.00	1.02	2.53
010000	1	0.56	0.11	1.29	0.35	1.94
001110	2	1.19	1.33	0.84	0.85	0.88

Table 7 (continued)

	Marginal Heterogeneity					
Rating Pattern	Observed Count	G	Gc	GP	GPc	GHeP
001100	2	2.27	1.58	1.54	0.78	2.05
000110	1	2.03	1.43	1.34	0.69	1.77
000101	1	1.80	1.28	1.16	0.61	1.58
000100	6	3.87	1.70	8.64	6.75	5.99
000001	1	0.36	0.07	0.75	0.21	1.25
000000	29	25.25	29.00	24.07	29.00	24.96

Table 7 also summarizes the predicted counts for the G agreement, Gc agreement, GP agreement, GPc agreement and GHeP agreement models considered under the assumption of marginal heterogeneity fitted to the intestinal biopsy data. Under the G model, 25.25 biopsies are predicted to be rated as having the absence of mucosecretion diminution and 4.75 biopsies are predicted to be rated as having the presence of mucosecretion diminution. In contrast to the G model under homogeneity, the remaining 38 ratings are *not* equally dispersed across the remaining 62 possible rating patterns. Instead, each rater's propensity to rate "absence" or "presence", is incorporated into how the counts are dispersed across the remaining 62 possible rating patterns. The GPc model, the best fitting model under the assumption of marginal heterogeneity, yields predicted cell counts that are closer to the observed cell counts for each of the 64 rating patterns.

This data set was the motivating example for my work and will be discussed further when the design of the simulation study is described in Chapter 3. I focus on log-linear models that categorize agreement homogeneous across categories (e.g., the G, GP, and GHeP models).

2.5 INTERPRETATION OF AGREEMENT PARAMETERS FOR A 2-CATEGORY OUTCOME

Rogel et al. (1998) interpret the agreement parameters for a 2-category outcome in the context of whether the type of agreement (global, partial and/or homogeneous or heterogeneous across categories) differs from the agreement expected by chance. The interpretation of the global and homogeneous partial agreement parameters is as follows: from the G model described by equation 2.1.14, d_6 can be written as

$d_6 = \log m^{(G)}_{i_1i_2i_3i_4i_5i_6} - (m + l^{O_1}_{i_1} + l^{O_2}_{i_2} + l^{O_3}_{i_3} + l^{O_4}_{i_4} + l^{O_5}_{i_5} + l^{O_6}_{i_6})$ where $\log m^{(G)}_{i_1i_2i_3i_4i_5i_6}$ i s the log expected value of $x_{i_1i_2i_3i_4i_5i_6}$ representing global agreement in this model. Letting $\log m^{(I')}_{i_1i_2i_3i_4i_5i_6} = (m + l^{O_1}_{i_1} + l^{O_2}_{i_2} + l^{O_3}_{i_3} + l^{O_4}_{i_4} + l^{O_5}_{i_5} + l^{O_6}_{i_6})$, where $\log m^{(I')}_{i_1i_2i_3i_4i_5i_6}$ is the part of $m^{(G)}_{i_1i_2i_3i_4i_5i_6}$ expected by chance, $e^{d_6} = m^{(G)}_{i_1i_2i_3i_4i_5i_6} / m^{(I')}_{i_1i_2i_3i_4i_5i_6}$ for i= 0, 1. If d_6 >0, then e^{d_6} >1 so $m^{(G)}_{i_1i_2i_3i_4i_5i_6} > m^{(I')}_{i_1i_2i_3i_4i_5i_6}$, global agreement is greater than that expected by chance. If d_6 < 0, then e^{d_6} <1 so $m^{(G)}_{i_1i_2i_3i_4i_5i_6} < m^{(I')}_{i_1i_2i_3i_4i_5i_6}$ and global agreement is less than that expected by chance. Similarly, the homogeneous partial agreement parameter d_5 can be rewritten as $\log m^{(GP)}_{i_1i_2i_3i_4i_5i_6} - \log m^{(G')}_{i_1i_2i_3i_4i_5i_6}$ with $e^{d_5} = m^{(GP)}_{i_1i_2i_3i_4i_5i_6} / m^{(G')}_{i_1i_2i_3i_4i_5i_6}$ where $m^{(GP)}_{i_1i_2i_3i_4i_5i_6}$ is the expected value of $x_{i_1i_2i_3i_4i_5i_6}$ under the GHeP model (equation 2.1.16) and $m^{(G')}_{i_1i_2i_3i_4i_5i_6}$ is the part of $m^{(GP)}_{i_1i_2i_3i_4i_5i_6}$ explained by global agreement and chance. If d_5 >0, then e^{d_5} >1 so $m^{(GP)}_{i_1i_2i_3i_4i_5i_6} > m^{(G')}_{i_1i_2i_3i_4i_5i_6}$ and homogeneous partial agreement is greater than expected by chance accounting for global agreement. If d_5 < 0, then e^{d_5} <1 so $m^{(GP)}_{i_1i_2i_3i_4i_5i_6} < m^{(G')}_{i_1i_2i_3i_4i_5i_6}$ and homogeneous partial agreement is less than expected by chance accounting for global agreement.

Each of the heterogeneous partial agreement parameters in equation 2.1.18, $d_5^{G_{\overline{p}}}$ (p = 1, …, 6) can be rewritten as $\log m_{i_1 i_2 i_3 i_4 i_5 i_6}^{(GHeP)} / m_{i_1 i_2 i_3 i_4 i_5 i_6}^{(GP')}$ with $e^{d_5^{G_{\overline{p}}}} = m_{i_1 i_2 i_3 i_4 i_5 i_6}^{(GHeP)} / m_{i_1 i_2 i_3 i_4 i_5 i_6}^{(GP')}$ where $m_{i_1 i_2 i_3 i_4 i_5 i_6}^{(GHeP)}$ is the expected value of $x_{i_1 i_2 i_3 i_4 i_5 i_6}$ under the GHeP model and $m_{i_1 i_2 i_3 i_4 i_5 i_6}^{(G')}$ is the part of $m_{i_1 i_2 i_3 i_4 i_5 i_6}^{(GHeP)}$ explained by global agreement and chance in this model. If $d_5^{G_{\overline{p}}}$ >0, then $e^{d_5^{G_{\overline{p}}}}$ >1 so $m_{i_1 i_2 i_3 i_4 i_5 i_6}^{(GHeP)} > m_{i_1 i_2 i_3 i_4 i_5 i_6}^{(GP')}$, there is more agreement among five raters when rater p is excluded than is expected when partial agreement is assumed to be homogeneous. If $d_5^{G_{\overline{p}}} < 0$, then $e^{d_5^{G_{\overline{p}}}}$ <1 so $m_{i_1 i_2 i_3 i_4 i_5 i_6}^{(GHeP)} < m_{i_1 i_2 i_3 i_4 i_5 i_6}^{(GP')}$, and there is less agreement among the five raters when rater p is excluded than is expected when partial agreement is assumed to be homogeneous.

Rogel et al. (1998) also quantify the magnitude of agreement between two raters via the conditional odds ratio computed from a log-linear model of pair-wise agreement. Conditioning on the ratings of raters 3, 4, 5, and 6 (O^3, O^4, O^5, O^6), the $d_{i_1 i_2 i_3 i_4 i_5 i_6}$ term when assessing pair-wise agreement between raters 1 and 2 (O^1, O^2) is defined as:

$$d_{i_1 i_2 i_3 i_4 i_5 i_6} = \left[\mathrm{I}\, d_2^{O_1 O_2} \right] \text{ where I =1 if } i_1 = i_2,$$

$$= 0, \text{ otherwise}$$

As described by Rogel et al. (1998), conditioning on the ratings of raters 3, 4, 5, and 6 (O^3, O^4, O^5, O^6), the odds ratio of agreement for rater 1 and rater 2 is written as

$$t_{i_1 i_2 (i_3 i_4 i_5 i_6)} = \frac{m_{i_1 i_1 i_3 i_4 i_5 i_6} m_{i_2 i_2 i_3 i_4 i_5 i_6}}{m_{i_1 i_2 i_3 i_4 i_5 i_6} m_{i_1 i_2 i_3 i_4 i_5 i_6}}, \text{ where } \log m_{i_1 i_1 i_3 i_4 i_5 i_6} = m + d_2^{O_1 O_2}, \ \log m_{i_2 i_2 i_3 i_4 i_5 i_6} = m + d_2^{O_1 O_2},$$

and $\log m_{i_1 i_2 i_3 i_4 i_5 i_6} = m$. Therefore, $t_{i_1 i_2 (i_3 i_4 i_5 i_6)} = \dfrac{e^{d_2^{O_1 O_2}} e^{d_2^{O_1 O_2}}}{e^0 e^0} = e^{2 d_2^{O_1 O_2}}$.

If agreement is homogeneous among pairs of raters, $d_2^{O_iO_j} = d_2$ for all i and j (i, j = 1 to K and $i \neq j$), then given the rating of the other four raters, the odds that the first rater indicated the presence of the lesion rather than absence is estimated as $e^{2\hat{d}_2}$ times higher when the second rater rated presence rather than absence of the lesion. If the agreement is heterogeneous among rater pairs, then this odds ratio will vary by rater pair.

For the GP model with subgroups of five raters, if any four raters agree on the presence of the lesion, the odds that the fifth rater indicates 'presence' rather than 'absence' is estimated as $e^{(\hat{d}_6 - 2\hat{d}_5)}$ higher when the sixth rater indicated 'presence' rather than 'absence'. The odds ratio $t_{i_1 i_2 i_3 i_4 i_5 i_6}$ is written as $\dfrac{m_{i_1=1 i_2=1 i_3=1 i_4=1 i_5=1 i_6=1} m_{i_1=1 i_2=1 i_3=1 i_4=1 i_5=0 i_6=0}}{m_{i_1=1 i_2=1 i_3=1 i_4=1 i_5=0 i_6=1} m_{i_1=1 i_2=1 i_3=1 i_4=1 i_5=1 i_6=0}}$,

where $m_{i_1=1 i_2=1 i_3=1 i_4=1 i_5=1 i_6=1}$ = $\exp(m + d_6)$, $m_{i_1=1 i_2=1 i_3=1 i_4=1 i_5=0 i_6=0}$ = $\exp(m)$, $m_{i_1=1 i_2=1 i_3=1 i_4=1 i_5=0 i_6=1}$ = $\exp(m + d_5^{\bar{5}})$ and $m_{i_1=1 i_2=1 i_3=1 i_4=1 i_5=1 i_6=0}$ = $\exp(m + d_5^{\bar{6}})$. Therefore, $t_{i_1 i_2 i_3 i_4 i_5 i_6} = \exp(m + d_6$ $+ m - m - d_5^{\bar{5}} - m - d_5^{\bar{6}})$. Since it is the GP model, and partial agreement is homogeneous, $t_{i_1 i_2 i_3 i_4 i_5 i_6} = \exp(d_6 - 2d_5)$.

For the GHeP model under the assumption of marginal homogeneity, the partial agreement is not homogeneous and $t_{i_1 i_2 i_3 i_4 i_5 i_6} = \exp(d_6 - d_5^{\bar{5}} - d_5^{\bar{6}})$. Using the estimates from Table 6, the odds that Rater 5 indicates 'presence' rather than 'absence' is estimated as ($\exp^{(3.58-0.87-1.27)}$ =) 4.2 times higher when Rater 6 indicates 'presence' rather than 'absence'.

2.6 INFERENCE

I review inference in the context of a family of hypotheses, the collection of hypotheses that are of interest for a dataset. I outline procedures to control Type I error when multiple hypotheses are tested.

2.6.1. Type One Error

Testing multiple hypotheses inflates the Type I error rate, defined as the probability of rejecting the null hypothesis when the null hypothesis is true. When the family (e.g., collection) of hypotheses includes more than one hypothesis, two kinds of Type I error rates are often considered: the comparison-wise error rate (CWE) and the family-wise (experiment-wise) error rate (FWE) (Klockars and Sax, 1986; Shaffer, 1995). The CWE is the probability of a Type I error occurring for a single hypothesis. The FWE is the probability that at least one hypothesis in the family of hypotheses is falsely rejected. When a family of hypotheses involves only one hypothesis, the CWE equals the FWE. When more than one hypothesis is tested, each at the same α level, the FWE is greater than the nominal level α. Consequently, the FWE needs to be controlled at the desired pre-specified level ?. Several multiple comparison procedures control the FWE at α by adjusting the CWE of each hypothesis tested. Multiple comparison procedures include those categorized as one-step (simultaneous inference) or step-wise (sequentially rejective) procedures (Shaffer, 1995; Ludbrook 1998). Each of these multiple comparison procedures makes some adjustment to the p-value of each comparison. I denote the p-value uncorrected for the number of comparison made by p_u, and the p-value corrected for the number of comparison made by one of the multiple comparison procedures by p_c .

2.7 MULTIPLE COMPARISON PROCEDURES

2.7.1. One-Step Procedures

One-step procedures, such as the Bonferroni and Sidak Inequalities, apply the same correction to the p-value of each tested hypothesis in the family. The rationale behind the Bonferroni and Sidak Inequalities is the following: for a family of g null hypotheses ($H_1,...,H_g$), let C be the event that at least one of the g comparisons is statistically significant under the null hypothesis. The goal is to maintain the probability of C, Pr(C), at γ, the desired a priori specified FWE. To determine at what significance level α* each of the g comparisons should be conducted, note that $P\left(\overline{C}\right)$, the probability that none of the g comparisons is statistically significant under the null hypothesis, has to equal 1-γ. If the hypotheses are independent, then $P\left(\overline{C}\right)=\left(1-a^*\right)^g=1-g$. Solving for a^* yields the Sidak inequality adjustment, a comparison-wise error rate of $a^*=1-\sqrt[g]{(1-g)}$ (Sidak 1967; Shaffer, 1986). If a^* is small, then 1- γ can be approximated by 1-gα*. Solving for a^* in this approximation yields the Bonferroni correction, α* = γ/g (Rosner, 1995). Testing each hypothesis at a comparison-wise error rate α* keeps the FWE at the pre-specified level γ. For the Bonferroni procedure, p_c is obtained by multiplying p_u by g , (i.e. p_c = p_u *g). The Sidak Inequality corrected p-value is p_c = $p_u \times [1-(1-p_u)^g]$.

Corrected p-values less than the apriori specified γ level are considered to be statistically significant.

2.7.2. Holm's Step-Down Procedure

The Bonferroni and Sidak corrections are conservative procedures for controlling FWE when the test statistics of the hypotheses are correlated. Increased power can be obtained by using Holm's step-down procedure (Holm, 1979) because the critical levels are larger than γ/g or $1-\sqrt[g]{(1-g)}$. The Holm's procedure for testing g hypotheses in terms of both the Bonferroni and Sidak corrections is:

(1) Rank order the uncorrected p-values for the g tested hypotheses in ascending order.

(2) Calculate p_c for the smallest p-value, $p_c = p_u * g$ for the Bonferroni inequality or $p_c = 1-(1- p_u)^g$ for the Sidak inequality.

(3) Calculate the p_c for the next smallest p-value as $p_c = (g-1) * p_u$ for the Bonferroni inequality or $p_c = 1-(1- p_u)^{g-1}$ for the Sidak inequality.

(4) Continue this step-wise procedure until the corrected p-value exceeds α, or all p-values have been corrected. Reject the null hypotheses associated with adjusted p-values less than γ and fail to reject the null hypotheses associated with adjusted p-values that exceed γ.

The Holm's procedure is based upon the closure method (closed testing procedure) proposed by Marcus et al. (1976) and the union intersection principle. If a hypothesis, H_b, can be expressed as the intersection of a finite family of g hypotheses, $H_b = H_{b_1} \cap H_{b_2} \cap ... \cap H_{b_q}$, then the rejection region of H_b is the union of the rejection regions of $H_{b_1}, H_{b_2}, ..., H_{b_k}$. The closure method requires that the intersection of the *k* hypotheses does not include the null set. Provided that a α level test is available for each $H_{b_i}, i =$ 1 to g, the closed testing procedure rejects any hypothesis *Hb* if and only if all $H_{b_i}, i =$ 1 to g, are rejected by the associated α level test. As such, the closed testing procedure controls the FWE at level ?. For example, let there be three raters, and let $H_b : d_2^{\bar{1}} = d_2^{\bar{2}} = d_2^{\bar{3}} = constant$ (i.e. the heterogeneous partial agreement parameters are homogeneous). Hypothesis *Hb*, can be expressed as the intersection of $H_{b_1} \cap H_{b_2} \cap H_{b_3}$, where $H_{b_1} : d_2^{\bar{1}} = d_2^{\bar{2}}$, $H_{b_2} : d_2^{\bar{1}} = d_2^{\bar{3}}$, and $H_{b_3} : d_2^{\bar{2}} = d_2^{\bar{3}}$. In addition, the H_{b_i} can be tested in order of the largest test statistic to the smallest test statistic. For example, if H_{b_2} has the largest test statistic, and H_{b_1} has the smallest test statistic, the *Hb* can be tested in the order of H_{b_2}, H_{b_3} and H_{b_1}. Ordering the test statistics from largest to smallest yields the order statistics of the corresponding p-values. The hypothesis corresponding to the smallest p-value (e.g., largest test statistic) is tested first as described above (here, 0.05/3). If it is rejected, the hypothesis corresponding to the second smallest p-value (H_{b_3}) is then tested (0.05/2). This procedure continues until one of the hypotheses is not rejected or all the hypotheses in the family have been rejected. The Holm's procedure controls the FWE at level γ.

Table 8 shows the corrected critical values for the 15 pair-wise comparisons when there are 6 raters for each of the multiple comparison procedures considered and the overall family-wise Type I error rate is ? = 0.05. The corrected critical p-values are $0.00\overline{3}$ and 0.0034 for the Bonferroni and Sidak adjustments, respectively. The corrected critical p-values for the first tested hypothesis are $0.00\overline{3}$ and 0.0034 for the Holm's-Bonferroni and Holm's-Sidak adjustments, respectively, and increase in magnitude as each preceding hypothesis is found significant.

The number of raters in the inter-rater agreement study dictates the number of pair-wise comparisons made. As the number of pair-wise comparisons increases, the value of the corrected critical p-value decreases, making it more difficult to reject the hypothesis. If, for example, there are sampling zeros for the rating patterns representing partial agreement for two raters, then pair-wise comparisons are made among the remaining 4 raters. For illustrative purposes, Table 9 summarizes the corrected critical p-values for the 6 pair-wise comparisons when there are 4 raters for each of the multiple comparison procedures considered. Note that the initial corrected critical p-value is larger compared to that shown in Table 8 for 6 raters, and when the Holm's Bonferroni or Holm's Sidak procedure is used, the increase in the magnitude of the next corrected critical p-value is bigger when fewer comparisons are made (Table 9).

Table 8. Value of the corrected critical p-values for the four multiple comparison procedures considered for the 15 pair-wise comparisons when there are 6 raters and the overall family-wise Type I error rate is gamma = 0.05.

Unconditional Pair-wise Comparisons *			**Conditional Pair-wise Comparisons** **		
Pair-wise Comparison, Hypothesis Tested	Bonferroni	Sidak	Pair-wise Comparison, Hypothesis Tested	Holm's – Bonferroni	Holm's - Sidak
$H_1: d_5^{\bar{1}}$ vs. $d_5^{\bar{2}}$	0.0033	0.003414	H_1 (H_1, p-w comparison with smallest p-value)	0.0033	0.0034
$H_2: d_5^{\bar{1}}$ vs. $d_5^{\bar{3}}$	0.0033	0.0034	$H_2 \mid H_{1\ significant}$ (H_2, p-w comparison with 2nd smallest p-value)	0.0035	0.0036
$H_3: d_5^{\bar{1}}$ vs. $d_5^{\bar{4}}$	0.0033	0.0034	$H_3 \mid H_{2\ significant}$ (H_3, p-w comparison with 3rd smallest p-value)	0.0038	0.0039
$H_4: d_5^{\bar{1}}$ vs. $d_5^{\bar{5}}$	0.0033	0.0034	$H_4 \mid H_{3\ significant}$ (H_4, p-w comparison with 4th smallest p-value)	0.0041	0.0042
$H_5: d_5^{\bar{1}}$ vs. $d_5^{\bar{6}}$	0.0033	0.0034	$H_5 \mid H_{4\ significant}$ (H_5, p-w comparison with 5th smallest p-value)	0.0045	0.0046
$H_6: d_5^{\bar{2}}$ vs. $d_5^{\bar{3}}$	0.0033	0.0034	$H_6 \mid H_{5\ significant}$ (H_6, p-w comparison with 6th smallest p-value)	0.005	0.0051
$H_7: d_5^{\bar{2}}$ vs. $d_5^{\bar{4}}$	0.0033	0.0034	$H_7 \mid H_{6\ significant}$ (H_7, p-w comparison with 7th smallest p-value)	0.0055	0.0056
$H_8: d_5^{\bar{2}}$ vs. $d_5^{\bar{5}}$	0.0033	0.0034	$H_8 \mid H_{7\ significant}$ (H_8, p-w comparison with 8th smallest p-value)	0.0062	0.0063

Table 8 (continued)

Unconditional Pair-wise Comparisons *			**Conditional Pair-wise Comparisons** **		
Pair-wise Comparison, Hypothesis Tested	Bonferroni	Sidak	Pair-wise Comparison, Hypothesis Tested	Holm's – Bonferroni	Holm's - Sidak
$H_9 : \bar{d}_5^{2}$ vs. $\bar{d}_5^{6}$	0.0033	0.0034	$H_9 \mid H_{8\ significant}$ (H_9, p-w comparison with 7th largest p-value)	0.0071	0.0073
$H_{10} : \bar{d}_5^{3}$ vs. $\bar{d}_5^{4}$	0.0033	0.0034	$H_{10} \mid H_{9\ significant}$ (H_{10}, p-w comparison with 6th largest p-value)	0.008333	0.0085
$H_{11} : \bar{d}_5^{3}$ vs. $\bar{d}_5^{5}$	0.0033	0.0034	$H_{11} \mid H_{10\ significant}$ (H_{11},p-w comparison with 5th largest p-value)	0.01	0.0102
$H_{12} : \bar{d}_5^{3}$ vs. $\bar{d}_5^{6}$	0.0033	0.0034	$H_{12} \mid H_{11\ significant}$ (H_{12}, p-w comparison with 4th largest p-value)	0.0125	0.0127
$H_{13} : \bar{d}_5^{4}$ vs. $\bar{d}_5^{5}$	0.0033	0.0034	$H_{13} \mid H_{12\ significant}$ (H_{13},p-w comparison with 3rd largest p-value)	0.016667	0.0169
$H_{14} : \bar{d}_5^{4}$ vs. $\bar{d}_5^{6}$	0.0033	0.0034	$H_{14} \mid H_{13\ significant}$ (H_{14}, p-w comparison with 2nd largest p-value)	0.025	0.0253
$H_{15} : \bar{d}_5^{5}$ vs. $\bar{d}_5^{6}$	0.0033	0.0034	$H_{15} \mid H_{14\ significant}$ (H_{15}, p-w comparison with largest p-value)	0.05	0.05

* Hypotheses of pair-wise comparisons are not ordered when using the Bonferroni or Sidak procedures.

** Hypotheses of pair-wise comparisons for the Holm's-Bonferroni or Holm's –Sidak procedures are ordered from the smallest to largest p-value.

Table 9. Value of the corrected critical p-values for the four multiple comparison procedures considered for the 15 pair-wise comparisons when there are 4 raters and the overall family-wise Type I error rate is gamma = 0.05.

Pair-wise Comparison, Hypothesis* Tested	Bonferroni	Sidak	Pair-wise Comparison, Hypothesis** Tested	Holm's – Bonferroni	Holm's - Sidak
$H_1 : d_5^{\bar{1}}$ vs. $d_5^{\bar{2}}$	0.0083	0.008512	H_1 (H_1, p-w comparison with smallest p-value)	0.0083	0.0085
$H_2 : d_5^{\bar{1}}$ vs. $d_5^{\bar{3}}$	0.0083	0.008512	$H_2 \mid H_{1\ significant}$ (H_2, p-w comparison with 2nd smallest p-value)	0.01	0.0102
$H_3 : d_5^{\bar{1}}$ vs. $d_5^{\bar{4}}$	0.0083	0.008512	$H_3 \mid H_{2\ significant}$ (H_3, p-w comparison with 3rd smallest p-value)	0.0125	0.0127
$H_4 : d_5^{\bar{2}}$ vs. $d_5^{\bar{3}}$	0.0083	0.008512	$H_4 \mid H_{3\ significant}$ (H_4, p-w comparison with 3rd largest p-value)	0.0166	0.0169
$H_5 : d_5^{\bar{2}}$ vs. $d_5^{\bar{4}}$	0.0083	0.008512	$H_5 \mid H_{4\ significant}$ (H_5, p-w comparison with 2nd largest p-value)	0.025	0.0253
$H_6 : d_5^{\bar{3}}$ vs. $d_5^{\bar{4}}$	0.0083	0.008512	$H_6 \mid H_{5\ significant}$ (H_6, p-w comparison with largest p-value)	0.05	0.05

* Hypotheses of pair-wise comparisons are not ordered when using the Bonferroni or Sidak procedures.

** Hypotheses of pair-wise comparisons for the Holm's-Bonferroni or Holm's –Sidak procedures are ordered from the smallest to largest p-value.

3. METHODS

The focus of this work is to formalize inferential approaches that can be used to test the assumption of rater exchangeability and identify an atypical rater in the framework of Rogel et al.'s log-linear models. I propose an unconditional approach to test the assumption of rater exchangeability and identify an atypical rater, based on fitting the GHeP model directly (without using a model selection process). The Type I error of the approach when raters are homogeneous or heterogeneous with respect to their marginal distributions and the power of this approach to identify a single atypical rater with rater sub-groups of size K-1 were assessed via a simulation study. These data were simulated from scenarios with known underlying structure of agreement. I also compared alternative adjustments for the multiple comparison problem (e.g., Bonferroni, Sidak, Holm's Step-down procedures (Bonferroni and Sidak adjustments). This chapter ends with a description of the simulation study.

3.1. INFERENTIAL APPROACH

The inferential approach involves:

(1) Fitting the heterogeneous partial agreement log-linear model to the data,

(2) Performing pair-wise comparisons of the *K* partial agreement parameters, $\underset{\sim}{d}{}^{\bar{i}}_{K-1}$, and adjusting the p-values for the number of multiple comparisons performed, and

(3) Identifying any $d_{K-1}^{\bar{i}}$ agreement parameters that are involved in statistically significant pair-wise comparisons.

To identify an ‘atypical rater’, I based statistical inference on the heterogeneous partial agreement parameters of the GHeP model. Each partial agreement parameter, $\hat{d}_{K-1}^{G_{\bar{i}}}$, i = 1 to K, represents the level of agreement among the subgroup of K-1 raters when rater i is not included in the rater subgroup. If the level of agreement among rater subgroups differs significantly by which rater is excluded, then at least one rater is identified as atypical (e.g., the assumption of rater exchangeability does not hold). Defined in terms of the heterogeneous partial agreement parameters the null hypothesis is H_0: $d_{K-1}^{G_{\bar{1}}} = d_{K-1}^{G_{\bar{2}}} = ... = d_{K-1}^{G_{\bar{K}}} = d_{K-1}^{G_{\bar{i}}}$ vs. H_A: at least one $d_{K-1}^{G_{\bar{i}}} \neq d_{K-1}^{G_{\bar{j}}}$, $i \neq j$ where i = 1 to K. If the null hypothesis is rejected, then the K partial agreement parameters are not homogeneous, prompting the question “Which partial agreement parameter is statistically significantly different from the others?” The magnitude of each estimated partial agreement coefficient corresponds to each rater’s non-chance contribution to five-way agreement after accounting for global agreement. Under the assumption of marginal homogeneity, an atypical rater’s non-chance contribution to five-way agreement after accounting for global agreement is higher than that of a rater who is not atypical and the atypical rater’s partial agreement parameter estimate would be significantly *larger* in magnitude relative to the other heterogeneous partial agreement parameter estimates. Under the assumption of marginal heterogeneity, the heterogeneous partial agreement parameter estimate for an atypical rater may not differ in a predictable way from the remaining heterogeneous partial agreement parameter estimate because it reflects only disagreement that is not explained by the atypical rater's marginal distribution. An alternative strategy would be to examine differences in the lambda parameters in the model. This work focused on the pair-wise comparisons of the heterogeneous partial agreement parameters directly. I investigated whether hypothesis testing involving the *K* heterogeneous partial agreement parameters, adjusted for multiple comparisons, would correctly identify which rater, if any, is atypical. The analysis was performed assuming marginal homogeneity for scenarios simulated under the assumption of marginal homogeneity. For scenarios simulated under the assumption of marginal heterogeneity, the analysis was performed twice, under the assumptions of both marginal homogeneity and marginal heterogeneity.

The GHeP model was fit to the data without prior model selection. Pair-wise comparisons of the *K* partial agreement parameters, $\underset{\sim}{d}_{K-1}^{\bar{i}}$, were made using the Bonferroni and Sidak Inequalities and the Holms-Bonferroni and Holms-Sidak procedures, as described in Chapter 2. These partial agreement parameters partition the overall 5-way agreement into components attributable to each rater. The premise is that, in the presence of an atypical rater, at least one heterogeneous partial agreement parameter would differ from at least one of the remaining *K*-1 partial agreement parameters, controlling for the assumed marginal structure. The pair-wise comparisons of the *K* partial agreement parameters constitute a family of hypotheses where g = *K*(*K*-1)/2.

These partial agreement parameters are asymptotically multivariate normal with mean $\underset{\sim}{d}_{K-1}^{\bar{i}}$ and variance-covariance Σ, asymptotically $\underset{\sim}{\hat{d}}_{K-1}^{G_{\bar{i}}} \sim \mathrm{MVN}(\underset{\sim}{d}_{K-1}^{G_{\bar{i}}}, \mathrm{S})$. The pair-wise comparisons can be conducted using Z statistics for the appropriate linear combinations of the delta parameters (Wald test statistics), with adjustment for the number of comparisons being made.

3.2. SIMULATION STUDY

3.2.1. Objectives

The primary objectives of the simulation study were to assess the level (probability of Type I error) and the power of the proposed approach to detect an atypical rater in the context of several scenarios motivated by the intestinal biopsy rating study. Five simulation scenarios were considered. I considered the proposed approach under the assumption that each of five models (G, GP, GHeP-rog, GHeP-atyp4a, and GHeP-atyp4b) was correct under the assumption of marginal homogeneity, and again, under the assumption of marginal heterogeneity.

For scenarios simulated assuming marginal homogeneity, hypothesis testing was conducted under the assumption of marginal homogeneity. For scenarios simulated under the assumption of marginal heterogeneity, the hypothesis testing was conducted twice, under the assumption of marginal homogeneity and under the assumption of marginal heterogeneity (Table 10).

Table 10. Summary of the Properties Assessed By the Analytic Approaches Used for Each Simulation Model

Simulation Model	**Simulation Assumption of Marginals**		
	Homogeneity	Heterogeneity	
	Analytic Assumption of Marginals	**Analytic Assumption of Marginals**	
	Homogeneity	Homogeneity	Heterogeneity
G	Type I error	Type I error	Type I error
GP	Type I error	Type I error	Type I error
GHeP-rog	Power	Power	Power
GHeP-atyp4a	Power	Power	Power
GHeP-atyp4b	Power	Power	Power

3.2.2. Design

Monte Carlo simulation was used to generate 1,000 simulations for each of the five models shown in Table 10 under the assumption of marginal homogeneity and marginal heterogeneity. One thousand simulations provide a 95% confidence interval half-width of 0.01 for the estimated level of 0.05 and a maximum half-width of 0.03 for the estimated power assuming the maximum binomial variance (when p=0.5).

A simulation consisted of generating rating data, the counts for each cell of the 2^6 contingency table. Therefore, one thousand 2^6 contingency tables were generated for each model under the assumption of marginal homogeneity, and under the assumption of marginal heterogeneity. The agreement structure within a given 2^6 contingency table was the agreement structure defined by the log-linear model that was used to generate the rating data.

Rating data for the scenarios involving the G, GP and GHeP-rog models were generated using as "true" values the parameter estimates obtained by fitting the corresponding model to the intestinal biopsy data. Each model was fitted to the published data to get realistic values for the simulations. Rating data for the G, GP, and GHeP-rog simulation scenarios were constructed in the following manner:

1. Fit the hypothesized model to the mucosecretion diminution data.
2. Capture the estimates of the parameters and variance-covariance matrix for the model fit in Step 1.
3. Randomly generate 1,000 realizations of each parameter in the model using the SAS % MVN macro (SAS Institute Inc; http://ftp.sas.com/techsup/download/stat/mvn.html) using the estimates from Step 2 as input parameters.
4. Generate counts for the 2^6 contingency table by randomly sampling from a Poisson distribution with mean equal to the exponentiated sum of the coefficients corresponding to the covariate pattern of each of the 64 possible rating patterns.
5. Repeat Step #4 for the 1,000 realizations generated in Step #3.

When fitting the models to the intestinal biopsy data (Rogel et al. 1998), the 'sum-to-zero' constraint was used for the rater effects (see Appendix A) and indicator variables (see Appendix B) were used for the agreement parameters, as in Rogel et al. (1998). Stata, version 8.2, software was used to fit the models. The estimates of the parameters for each scenario are summarized in Table 11 (see Appendix C for Stata code and parameter estimate and variance-covariance matrix output).

Table 11. Marginal and Agreement Parameter Estimates for Five Possible Agreement Models Fitted to the Mucosecretion Diminution Intestinal Biopsy Data of Rogel et al. (1998) Assuming (a) Marginal Homogeneity and (b) Marginal Heterogeneity

	Assumed Marginal Homogeneity				
Parameter	G	GP	GHeP-rog	GHeP-atyp4a	GHeP-atyp4b
d_6	3.20	3.58	3.58	3.58	3.58
d_5	- -	1.22	- -	- -	- -
$d_5^{\bar{1}}$	- -	- -	0.87	0.96	0.96
$d_5^{\bar{2}}$	- -	- -	1.27	0.96	0.96
$d_5^{\bar{3}}$	- -	- -	0.17	0.96	0.96
$d_5^{\bar{4}}$	- -	- -	1.96	1.96	2.21
$d_5^{\bar{5}}$	- -	- -	0.87	0.96	0.96
$d_5^{\bar{6}}$	- -	- -	1.27	0.96	0.96
µ	-0.49	-0.87	-0.87	-0.87	-0.87

	Assumed Marginal Heterogeneity				
Parameter	G	GP	GHeP-rog	GHeP-atyp4a	GHeP-atyp4b
l^{O_1}	-0.51	-0.52	-0.64	-0.64	-0.64
l^{O_2}	-0.16	-0.10	-0.25	-0.24	-0.24
l^{O_3}	-0.26	-0.24	-0.35	-0.36	-0.36
l^{O_4}	0.80	0.84	1.35	1.35	1.35
l^{O_5}	-0.32	-0.30	-0.42	-0.42	-0.42
l^{O_6}	-0.38	-0.37	-0.48	-0.49	-0.49
d_6	3.47	3.96	4.50	4.49	4.49
d_5	- -	1.25	- -	- -	- -
$d_5^{\bar{1}}$	- -	- -	1.96	2.13	2.13
$d_5^{\bar{2}}$	- -	- -	2.44	2.13	2.13
$d_5^{\bar{3}}$	- -	- -	1.38	2.13	2.13
$d_5^{\bar{4}}$	- -	- -	0.36	0.37	**2.21**
$d_5^{\bar{5}}$	- -	- -	2.08	2.13	2.13
$d_5^{\bar{6}}$	- -	- -	2.47	2.13	2.13
µ	-1.08	-1.48	-2.08	-2.08	-2.08

The GHeP-rog estimates shown in Table 11 are the same as those shown in Table 6. The interpretation of the $d_5^{\bar{i}}$ and l^{O_i}, $i = 1$ to 6, was provided in section 2.4. Rating data for the GHeP-atyp4a scenario assuming marginal homogeneity was simulated using parameter estimates obtained by fitting a comparable model to the intestinal biopsy data. In the GHeP model, five of the six heterogeneous partial agreement parameters were constrained to be equal and the sixth was allowed to differ. Specifically, the heterogeneous partial agreement parameter for Rater 4 was allowed to differ, yielding estimates of 3.58 for d_6, -0.87 for µ, 0.96 for $\hat{d}_5^{\bar{i}}$, i =1, 2, 3, 5, and 6 and 1.96 for $\hat{d}_5^{\bar{4}}$ (fifth column of the first half of Table 11). Rating data for the GHeP-atyp4b scenario under the assumption of marginal homogeneity was created by using the same parameter estimates from the GHeP-atyp4a except the magnitude of $d_5^{\bar{4}}$ was increased to 2.21 (Table 11, sixth column). An increase of 0.25, from 1.96 to 2.21, represents a two-fold increase on a log scale.

Rating data for the GHeP-atyp4a scenario assuming marginal heterogeneity was created by using parameter estimates obtained by fitting a comparable GHeP model to the intestinal biopsy data that constrained five of the six heterogeneous partial agreement parameters to be equal and allowed the parameter for Rater 4 to differ. Estimates for $\hat{l}^{O_1}$ (i = 1 to 6) ranged from -0.64 to 1.35. The parameter estimate of l^{O_4} is relatively large under the G and GP models as well as the GHeP models. The estimate of $\hat{d}_5^{\bar{i}}$ (i =1, 2, 3, 5, and 6) was 2.13 and 0.37 for $\hat{d}_5^{\bar{4}}$, with $\hat{m}$ = -2.08 (fifth column of the second half of Table 11). Rating data for the GHeP-atyp4b scenario under the assumption of marginal heterogeneity was created by using the comparable parameter estimates from the GHeP-atyp4a scenario except the magnitude of $\hat{d}_{5}^{\bar{4}}$ was changed from 0.37 to 2.21.

These parameter estimates and the corresponding variance-covariance matrices were used with the SAS macro **MVN** (Appendix D) to generate 1,000 realizations of each parameter in the hypothesized model (Step #3). The **MVN** macro generates multivariate normal data using the Cholesky decomposition of the variance-covariance matrix, an approach commonly used to simulate multivariate normally distributed data (Kennedy and Gentle, 1980). The random number generator in the macro uses the time from the computer's internal clock as the seed for each run.

For each of the 1,000 realizations, count data were generated for the 2^6 contingency table by randomly sampling from a Poisson distribution with mean equal to the exponentiated sum of the coefficients corresponding to the covariate pattern of each of the 64 possible rating patterns (Appendix E). For the GHeP-atyp4a model under marginal homogeneity, the portion of SAS code that generates the count data was as follows:

```
logm=mu4+e6*b1+e5sub*b2+e5m4*b3;

cntN&index=exp(logm);

smcnt&index = ranpoi(0,cntN&index);
```

where variables **e6**, **e5sub** and **e5m4** hold the value '0' or '1' defined by the rating pattern and **mu4, b1**, **b2**, and **b3** were one set of realized parameter estimates for the grand mean, global agreement, five-way agreement when raters 1, 2, 3, 5, or 6 are the discrepant rater, and five-way agreement when rater 4 is the discrepant rater, respectively. For example, for rating patterns (000100) and (111011) variable **e6** equals zero because neither rating pattern represents global agreement, variable **e5sub** equals zero because the five-way agreement represented by these two ratings patterns was not the result of raters 1, 2, 3, 5 or 6 being the discrepant rater, and variable **e5m4** equals one because rater 4 was the discrepant rater and the pattern represented five-way agreement. If one set of realized parameter estimates was (-1.01, 3.64, 1.18, 2.37) for the variables **mu4, b1**, **b2**, and **b3**, respectively, the value of variable **logm** for rating patterns (000100) and (111011) is

$$\log m_{000100} = -1.01 + (d_6 = 0)*3.64 + (d_{5sub} = 0)*1.18 + (d_5^{\bar{4}} = 1)*2.37 = -1.01 + 2.37 = 1.36$$

Count data for rating pattern (000100) was generated by sampling from a Poisson distribution with mean equal to 1.36. Count data for rating pattern (111011) was also generated by sampling from a Poisson distribution with mean equal to 1.36. The means of the two Poisson distributions are the same because the agreement is not assumed to vary by category of the response.

An example of simulated count data for each of the five models simulated assuming marginal homogeneity is provided in Table 12. The shaded rows highlight the rating patterns that represent global agreement or partial agreement. Note, the total number of the counts observed in a realized 2^6 contingency table, the sample size, is not fixed. For the five realized 2^6 contingency tables presented in Table 12, the sample size ranged from 58 to 95.

Table 12. One Set of Simulated Count Data for Each of the Five Scenarios Assuming Marginal Homogeneity

Rating Pattern	G Model	GP Model	GHeP-rog Model	GHeP-atyp4a Model	GHeP-atyp4b Model
000000	18	14	9	18	27
000001	0	1	2	0	0
000010	2	5	0	2	1
000011	1	0	0	1	0
000100	1	2	2	4	5
000101	1	0	0	0	0
000110	0	0	0	0	0
000111	1	0	1	0	2
001000	1	3	0	2	1
001001	3	2	0	0	0
001010	2	0	0	0	1
001011	0	0	1	0	0
001100	1	0	0	0	1
001101	1	0	0	1	1
001110	1	0	1	0	0
001111	1	0	0	0	1
010000	1	0	0	0	1
010001	2	0	0	0	1
010010	2	0	0	1	0
010011	2	0	1	1	0
010100	0	1	1	0	2
010101	0	0	1	0	2
010110	0	0	1	1	0
010111	0	0	0	1	0
011000	1	2	1	1	0
011001	2	2	1	0	0
011010	0	3	0	0	0
011011	1	0	0	0	0
011100	1	0	1	1	0
011101	0	0	0	0	0
011110	2	0	2	0	0
011111	0	2	1	0	0
100000	1	0	0	1	1
100001	0	0	0	2	0
100010	3	0	0	1	0
100011	0	1	0	0	1
100100	0	1	0	0	0
100101	1	0	2	0	0
100110	0	1	2	1	1
100111	0	0	1	1	1
101000	1	0	0	0	0
101001	2	0	1	0	0
101010	0	0	0	0	0
101011	0	0	2	0	0
101100	0	1	2	0	0
101101	0	1	0	0	1
101110	2	1	1	0	2
101111	1	2	1	2	1

Table 12 (continued)

Rating Pattern	G Model	GP Model	GHeP-rog Model	GHeP-atyp4a Model	GHeP-atyp4b Model
110000	1	0	1	0	1
110001	0	0	0	1	0
110010	0	1	0	1	1
110011	0	0	1	0	0
110100	1	0	1	0	2
110101	1	0	0	1	1
110110	1	0	0	0	0
110111	1	0	1	2	0
111000	0	0	0	0	1
111001	0	0	1	0	1
111010	1	0	0	0	1
111011	0	1	1	2	1
111100	0	0	0	0	0
111101	1	1	2	2	1
111110	1	2	2	1	0
111111	17	10	10	18	31
Sample Size	**84**	**60**	**58**	**70**	**95**

After the count data for the 1,000 2^6 contingency tables were generated for a given scenario, the analysis was conducted on each generated contingency table. When fitting each model to the data, parameters with sufficient statistics equal to zero had to be taken into account. My SAS program included code that ascertained which of the heterogeneous partial agreement parameters for a given contingency table had sufficient statistics equal to zero. The sufficient statistic of a heterogeneous partial agreement parameter was zero when both rating patterns representative of that parameter had counts of zero. For example, if rating patterns (010000) and (101111) both had zero counts (e.g., sampling zeros), the sufficient statistic of the heterogeneous partial agreement parameter for Rater 2, $d_5^{\bar{2}}$, was non-estimable, and was set to zero when the GHeP model was fitted to the data. There are 64 possible variations of the GHeP model when there are six raters and a binary outcome. A model number, 1 through 64, was assigned to each realized 2^6 contingency table. This model number was used as a data management tool to facilitate data processing when fitting the GHeP model to the data and when performing pair-wise comparisons of the GHeP parameters. The number of possible pair-wise comparisons of the GHeP parameters depends on the number of GHeP parameters with sufficient statistics not equal to zero. Table 13 enumerates each of the 64 possible GHeP models, summarizes the model number assigned to a given GHeP model, the number of possible pair-wise comparisons among the heterogeneous partial agreement parameters and the number of sampling zeros. When there are six raters, the number of possible pair-wise comparisons of the heterogeneous partial agreement parameters ranges from zero to 15.

In practice, in situations where an overall test is to be performed before individual comparisons, multiple comparison procedures generally are not used unless the overall test is statistically significant. However, to assess whether this strategy could identify pair-wise differences in the absence of a significant overall test (e.g., under circumstances when an investigator would not have planned to initially perform an overall test) while controlling for Type I error, I computed adjusted p-values regardless of the statistical significance of the overall test.

Table 13. Enumerated GHeP Models Having Heterogeneous Partial Agreement Parameters with Sufficient Statistics Equal to Zero and Its Number of Possible Pair-wise Comparisons

Model Number	**Heterogeneous Partial Agreement Parameter with Sufficient Statistic Equal to Zero**	**# of Pair-wise Comparisons (# Sampling Zeros)**
1	None of the six parameters	15 (0)
2	$d_5^{\overline{6}}$	10 (2)
3	$d_5^{\overline{5}}$	
4	$d_5^{\overline{4}}$	
5	$d_5^{\overline{3}}$	
6	$d_5^{\overline{2}}$	

Table 13 (continued)

Model Number	Heterogeneous Partial Agreement Parameter with Sufficient Statistic Equal to Zero	# of Pair-wise Comparisons (# Sampling Zeros)
7	$d_5^{\bar{1}}$	6 (4)
8	$d_5^{\bar{1}}, d_5^{\bar{2}}$	
9	$d_5^{\bar{1}}, d_5^{\bar{3}}$	
10	$d_5^{\bar{1}}, d_5^{\bar{4}}$	
11	$d_5^{\bar{1}}, d_5^{\bar{5}}$	
12	$d_5^{\bar{1}}, d_5^{\bar{6}}$	
13	$d_5^{\bar{2}}, d_5^{\bar{3}}$	
14	$d_5^{\bar{2}}, d_5^{\bar{4}}$	
15	$d_5^{\bar{2}}, d_5^{\bar{5}}$	
16	$d_5^{\bar{2}}, d_5^{\bar{6}}$	
17	$d_5^{\bar{3}}, d_5^{\bar{4}}$	
18	$d_5^{\bar{3}}, d_5^{\bar{5}}$	
19	$d_5^{\bar{3}}, d_5^{\bar{6}}$	
20	$d_5^{\bar{4}}, d_5^{\bar{5}}$	
21	$d_5^{\bar{4}}, d_5^{\bar{6}}$	
22	$d_5^{\bar{5}}, d_5^{\bar{6}}$	
23	$d_5^{\bar{1}}, d_5^{\bar{2}}, d_5^{\bar{3}}$	3 (6)
24	$d_5^{\bar{1}}, d_5^{\bar{2}}, d_5^{\bar{4}}$	
25	$d_5^{\bar{1}}, d_5^{\bar{2}}, d_5^{\bar{5}}$	
26	$d_5^{\bar{1}}, d_5^{\bar{2}}, d_5^{\bar{6}}$	
27	$d_5^{\bar{2}}, d_5^{\bar{3}}, d_5^{\bar{4}}$	
28	$d_5^{\bar{2}}, d_5^{\bar{3}}, d_5^{\bar{5}}$	
29	$d_5^{\bar{2}}, d_5^{\bar{3}}, d_5^{\bar{6}}$	
30	$d_5^{\bar{3}}, d_5^{\bar{4}}, d_5^{\bar{5}}$	
31	$d_5^{\bar{3}}, d_5^{\bar{4}}, d_5^{\bar{6}}$	
32	$d_5^{\bar{1}}, d_5^{\bar{3}}, d_5^{\bar{4}}$	
33	$d_5^{\bar{1}}, d_5^{\bar{3}}, d_5^{\bar{5}}$	
34	$d_5^{\bar{1}}, d_5^{\bar{3}}, d_5^{\bar{6}}$	

Table 13 (continued)

Model Number	**Heterogeneous Partial Agreement Parameter with Sufficient Statistic Equal to Zero**	**# of Pair-wise Comparisons (# Sampling Zeros)**
35	$d_5^{\bar{1}}, d_5^{\bar{4}}, d_5^{\bar{5}}$	3 (6)
36	$d_5^{\bar{1}}, d_5^{\bar{5}}, d_5^{\bar{6}}$	
37	$d_5^{\bar{1}}, d_5^{\bar{4}}, d_5^{\bar{6}}$	
38	$d_5^{\bar{2}}, d_5^{\bar{4}}, d_5^{\bar{5}}$	
39	$d_5^{\bar{2}}, d_5^{\bar{4}}, d_5^{\bar{6}}$	
40	$d_5^{\bar{4}}, d_5^{\bar{5}}, d_5^{\bar{6}}$	
41	$d_5^{\bar{3}}, d_5^{\bar{5}}, d_5^{\bar{6}}$	
42	$d_5^{\bar{2}}, d_5^{\bar{5}}\ d_5^{\bar{6}}$	
43	$d_5^{\bar{3}}, d_5^{\bar{4}}, d_5^{\bar{5}}, d_5^{\bar{6}}$	1 (8)
44	$d_5^{\bar{2}}, d_5^{\bar{4}}, d_5^{\bar{5}}, d_5^{\bar{6}}$	
45	$d_5^{\bar{2}}, d_5^{\bar{3}}, d_5^{\bar{5}}, d_5^{\bar{6}}$	
46	$d_5^{\bar{2}}, d_5^{\bar{3}}, d_5^{\bar{4}}, d_5^{\bar{6}}$	
47	$d_5^{\bar{2}}, d_5^{\bar{3}}, d_5^{\bar{4}}, d_5^{\bar{5}}$	
48	$d_5^{\bar{1}}, d_5^{\bar{4}}, d_5^{\bar{5}}, d_5^{\bar{6}}$	
49	$d_5^{\bar{1}}, d_5^{\bar{3}}, d_5^{\bar{5}}, d_5^{\bar{6}}$	
50	$d_5^{\bar{1}}, d_5^{\bar{3}}, d_5^{\bar{4}}, d_5^{\bar{6}}$	1 (8)
51	$d_5^{\bar{1}}, d_5^{\bar{3}}, d_5^{\bar{4}}, d_5^{\bar{5}}$	
52	$d_5^{\bar{1}}, d_5^{\bar{2}}, d_5^{\bar{5}}, d_5^{\bar{6}}$	
53	$d_5^{\bar{1}}, d_5^{\bar{2}}, d_5^{\bar{4}}, d_5^{\bar{6}}$	
54	$d_5^{\bar{1}}, d_5^{\bar{2}}, d_5^{\bar{4}}, d_5^{\bar{5}}$	
55	$d_5^{\bar{1}}, d_5^{\bar{2}}, d_5^{\bar{3}}, d_5^{\bar{6}}$	
56	$d_5^{\bar{1}}, d_5^{\bar{2}}, d_5^{\bar{3}}, d_5^{\bar{5}}$	
57	$d_5^{\bar{1}}, d_5^{\bar{2}}, d_5^{\bar{3}}, d_5^{\bar{4}}$	
58	$d_5^{\bar{2}}, d_5^{\bar{3}}, d_5^{\bar{4}}, d_5^{\bar{5}}, d_5^{\bar{6}}$	0* (10)
59	$d_5^{\bar{1}}, d_5^{\bar{3}}, d_5^{\bar{4}}, d_5^{\bar{5}}, d_5^{\bar{6}}$	
60	$d_5^{\bar{1}}, d_5^{\bar{2}}, d_5^{\bar{4}}, d_5^{\bar{5}}, d_5^{\bar{6}}$	
61	$d_5^{\bar{1}}, d_5^{\bar{2}}, d_5^{\bar{3}}, d_5^{\bar{5}}, d_5^{\bar{6}}$	
62	$d_5^{\bar{1}}, d_5^{\bar{2}}, d_5^{\bar{3}}, d_5^{\bar{4}}, d_5^{\bar{6}}$	

Table 13 (continued)

Model Number	Heterogeneous Partial Agreement Parameter with Sufficient Statistic Equal to Zero	# of Pair-wise Comparisons (# Sampling Zeros)
63	$d_5^{\bar{1}}$, $d_5^{\bar{2}}$, $d_5^{\bar{3}}$, $d_5^{\bar{4}}$, $d_5^{\bar{5}}$	0* (10)
64	All six parameters	0 (12)

* The hypothesis that $e5m_i = 0$ can be tested.

Models 58 through 63 have only one heterogeneous partial agreement parameter that is not constrained to be zero. Although this parameter can be tested, pair-wise comparisons are not possible. If the heterogeneous partial agreement parameter estimate is not significantly different from zero, the five-way agreement represented by the heterogeneous partial agreement parameter is not more than what would be expected by chance alone after accounting for global agreement (and marginal heterogeneity if assumed).

This inferential approach was implemented using the statistical software Stata 8.2 and SAS 8.2. I wrote programs to fit the five GHeP models, perform pair-wise comparisons, and compute the unadjusted p-value associated with each pair-wise comparison in Stata. SAS was used to compute the adjusted p-value values of each pair-wise comparison and for data management purposes in computing the following:

1. A summary of the mode, minimum and maximum sample size across the 1,000 simulations simulated from the G, GP, GHeP-rog, GHeP-atyp4a or GHeP-atyp4b model.
2. Descriptive statistics of the marginal percentages for different category specific agreement patterns and rater exclusion across the 1,000 simulations simulated from the G, GP, GHeP-rog, GHeP-atyp4a or GHeP-atyp4b model.

3. The distribution of the simulated 2^6 contingency tables having none, some or all of its heterogeneous partial agreement parameters with sufficient statistic equal to zero for each simulation model considered.

4. The frequency that each of the fifteen pair-wise comparisons was statistically significant. These statistics were summarized by multiple comparison procedure and number of pair-wise comparisons/ sampling zeros subsets. For simulations having one or more sampling zeros, an adjustment to the denominator was made when calculating the proportion. The number of simulations in which a given pair-wise comparison was not possible due to sampling zeros was subtracted from the total number of simulations considered.

5. Type I error is defined as the proportion of simulated tables generated under the G or GP model that had a significant pair-wise comparison involving any of the heterogeneous partial agreement parameters. An indicator variable, denoting at least one pair-wise comparison (any pair-wise comparison) was significant, was created for each simulation. Type I error was computed as the ratio of the number of simulated tables with at least one significant pair-wise comparison to 1,000 or (1000-*X*), where *X* is the number of simulations that have all$d_5^{\overline{i}}$, i=1 to 6, with sufficient statistic equal to zero.

6. Power is defined as the proportion of replications generated under the GHeP-rog, GHeP-atyp4a or GHeP-atyp4b models that had the designated atypical rater (Rater 4) identified as atypical by the multiple comparison procedures (i.e. at least of the following pair-wise comparisons was statistically significant $d_5^{\bar{1}}$ vs. $d_5^{\bar{4}}$, $d_5^{\bar{2}}$ vs. $d_5^{\bar{4}}$, $d_5^{\bar{3}}$ vs. $d_5^{\bar{4}}$, $d_5^{\bar{4}}$ vs. $d_5^{\bar{5}}$, or $d_5^{\bar{4}}$ vs. $d_5^{\bar{6}}$ for a given simulated 2^6 contingency table). The denominator was 1,000 or (1000-*X*), as appropriate. I summarized the proportion of simulated tables for which each possible number of the pair-wise comparisons involving $d_5^{\bar{4}}$ was significant.

7. The proportion of replications that identified a rater other than rater 4 as being the atypical rater was estimated. A rater other than rater 4 was considered identified as the atypical rater if one or more of the following pair-wise comparisons was statistically significant:$d_5^{\bar{1}}$ vs. $d_5^{\bar{2}}$, $d_5^{\bar{1}}$ vs. $d_5^{\bar{3}}$, $d_5^{\bar{1}}$ vs. $d_5^{\bar{5}}$, $d_5^{\bar{1}}$ vs. $d_5^{\bar{6}}$, $d_5^{\bar{2}}$ vs. $d_5^{\bar{3}}$, $d_5^{\bar{2}}$ vs. $d_5^{\bar{5}}$, $d_5^{\bar{2}}$ vs. $d_5^{\bar{6}}$, $d_5^{\bar{3}}$ vs. $d_5^{\bar{5}}$, $d_5^{\bar{3}}$ vs. $d_5^{\bar{6}}$ or $d_5^{\bar{5}}$ vs. $d_5^{\bar{6}}$ for a given simulated 2^6 contingency table. The denominator was 1,000 or (1000-*X*), as appropriate.

8. The overall unconditional probability of identifying an atypical rater was estimated by the proportion of pair-wise comparisons that were significant across the 1,000 simulated tables.

9. The overall conditional probability of identifying Rater 4 as the atypical rater was estimated by the proportion of pair-wise comparisons that identified Rater 4, given that at least one rater was identified as atypical.

4. RESULTS

The first section of this chapter illustrates the inferential approach using the intestinal biopsy rating data from Rogel et al. (1998). The second section summarizes the results from the simulation study described in Chapter 3.

4.1. ANALYSIS OF INTESTINAL BIOPSY RATING DATA

The mucosecretion diminution data and the observed rating patterns were described in section 2.4. My work required replicating some of the work done by Rogel et al. (1998). Reproducing results from the paper provided a way to validate my programs.

4.1.1. Results for the GHeP Model Assuming Marginal Homogeneity

The unconditional approach for identifying an atypical rater begins by fitting the GHeP model assuming marginal homogeneity to the data. For the intestinal biopsy data, the estimates of the rater effects, global agreement and the six partial agreement parameters are shown in Table 6.

Next, pair-wise comparisons of the heterogeneous partial agreement parameters with adjusted p-values were made. Table 14 summarizes the unadjusted and adjusted p-values using the Bonferroni, Sidak, Holm’s-Bonferroni and Holm’s-Sidak procedures for this GHeP model. The smallest unadjusted p-value was 0.10, which was comparing 5-way agreement excluding Rater 3 with 5-way agreement excluding Rater 4. These two raters had the most discrepant delta parameters (0.17 and 1.96, respectively) in Table 6. Adjusted p-values range from 0.78 to >0.99 for the Sidak and Holm’s-Sidak adjustments and were consistently >0.99 for the Bonferroni and Holm's-Bonferroni adjustments. None of the 15 pair-wise comparisons of the heterogeneous partial agreement parameters from this GHeP model was statistically significant.

Table 14. Unadjusted and Adjusted p-values for the Fifteen Pair-wise Comparisons of the Six Heterogeneous Partial Agreement Parameters from the GHeP Model Assuming Marginal Homogeneity.

	GHeP Model Fitted Assuming Marginal Homogeneity				
Comparison	Unadjusted	Bonferroni	Holm's – Bonferroni	Sidak	Holm's – Sidak
$d_5^{\bar{1}}$ vs. $d_5^{\bar{2}}$	0.66	> 0.99	> 0.99	> 0.99	0.99
$d_5^{\bar{1}}$ vs. $d_5^{\bar{3}}$	0.57	> 0.99	> 0.99	> 0.99	0.99
$d_5^{\bar{1}}$ vs. $d_5^{\bar{4}}$	0.17	> 0.99	> 0.99	0.95	0.94
$d_5^{\bar{1}}$ vs. $d_5^{\bar{5}}$	1.00	> 0.99	> 0.99	> 0.99	> 0.99
$d_5^{\bar{1}}$ vs. $d_5^{\bar{6}}$	0.66	> 0.99	> 0.99	> 0.99	0.99
$d_5^{\bar{2}}$ vs. $d_5^{\bar{3}}$	0.34	> 0.99	> 0.99	0.99	0.99
$d_5^{\bar{2}}$ vs. $d_5^{\bar{4}}$	0.34	> 0.99	> 0.99	0.99	0.99
$d_5^{\bar{2}}$ vs. $d_5^{\bar{5}}$	0.66	> 0.99	> 0.99	> 0.99	0.99
$d_5^{\bar{2}}$ vs. $d_5^{\bar{6}}$	> 0.99	> 0.99	> 0.99	> 0.99	> 0.99
$d_5^{\bar{3}}$ vs. $d_5^{\bar{4}}$	0.10	> 0.99	> 0.99	0.78	0.78
$d_5^{\bar{3}}$ vs. $d_5^{\bar{5}}$	0.57	> 0.99	> 0.99	> 0.99	0.99
$d_5^{\bar{3}}$ vs. $d_5^{\bar{6}}$	0.34	> 0.99	> 0.99	0.99	0.99
$d_5^{\bar{4}}$ vs. $d_5^{\bar{5}}$	0.17	> 0.99	> 0.99	0.95	0.94
$d_5^{\bar{4}}$ vs. $d_5^{\bar{6}}$	0.33	> 0.99	> 0.99	0.99	0.99
$d_5^{\bar{5}}$ vs. $d_5^{\bar{6}}$	0.66	> 0.99	> 0.99	> 0.99	0.99

4.1.2. Results for the GHeP Model Assuming Marginal Heterogeneity

Under the assumption of marginal heterogeneity, the each rater's overall prevalence of a positive rating for mucosecretion diminution is estimated. The largest lambda in Table 6, $l^{O_4} = 1.35,$ for Rater 4 corresponds to the largest marginal percentage for the presence of mucosecretion diminution (54.4%) in Table 5. Relatively large lambda estimates correspond to relatively large contributions to the fitted counts for positive ratings. In this example, Rater 4 has relatively more positive ratings than any of the other raters. This GHeP log-linear model for the expected cell counts is:

$$\log m_{i_1 i_2 i_3 i_4 i_5 i_6} = -2.08 - 0.65\lambda_1 - 0.25\lambda_2 - 0.35\lambda_3 + 1.35\lambda_4 - 0.42\lambda_5 - 0.48\lambda_6 + 4.5\delta_6 + 1.96\delta_6^{\bar{1}}$$
$$+ 2.44\delta_6^{\bar{2}} + 1.38\delta_6^{\bar{3}} + 0.36\delta_6^{\bar{4}} + 2.08\delta_6^{\bar{5}} + 2.47\delta_6^{\bar{6}}.$$

The interpretation of the heterogeneous partial agreement parameters changes when marginal heterogeneity is allowed. Under the assumption of marginal homogeneity, the largest heterogeneous partial agreement parameter corresponded to the rater who disagreed relatively more often than the other five raters when five-way agreement was considered. Under marginal heterogeneity, the largest heterogeneous partial agreement parameter corresponds to the rater who disagrees relatively more often than the other five raters when five-way agreement is considered and this disagreement is not accounted for by the rater's propensity to assign a particular rating. The estimated heterogeneous partial agreement parameter for Rater 4 is 1.96 under the assumption of marginal homogeneity and only 0.36 under the assumption of marginal heterogeneity (Table 11). In the GHeP marginal heterogeneity model, the more frequent occurrence of five-way agreement where Rater 4 is the discrepant rater may be attributable to Rater 4's higher propensity to rate the presence of mucosecretion diminution. While strategies based on pair-wise comparisons of the heterogeneous partial agreement parameters may identify Rater 4 as "different" it is not necessarily because the corresponding delta parameter is large.

Table 15 summarizes the unadjusted and adjusted p-values using the alternative multiple comparison procedures after fitting the GHeP model assuming marginal heterogeneity. The unadjusted comparisons indicate that five-way agreement excluding Rater 4 differs significantly from five-way agreement excluding either Rater 2 or Rater 6. Raters 2 and 6 have the largest $\hat{d}_5^{\bar{i}}$ parameters in Table 11. None of the fifteen adjusted pair-wise comparisons were statistically significant, indicating that the significant unadjusted differences could be attributable to Type I error. The adjusted p-values range from 0.46 to >0.99 for the Holm's- Bonferroni and Holm's-Sidak procedures, and from 0.62 to >0.99 for the Bonferroni and Sidak procedures.

Table 15. Unadjusted p-values and Four Adjusted p-values for the Fifteen Pair-wise Comparisons of the Six Heterogeneous Partial Agreement Parameter from the GHeP Assuming Marginal Heterogeneity.

	GHeP Model Fitted Assuming Marginal Heterogeneity				
Comparison	Unadjusted	Bonferroni	Holm's – Bonferroni	Sidak	Holm's - Sidak
$d_5^{\bar{1}}$ vs. $d_5^{\bar{2}}$	0.61	> 0.99	> 0.99	> 0.99	> 0.99
$d_5^{\bar{1}}$ vs. $d_5^{\bar{3}}$	0.64	> 0.99	> 0.99	> 0.99	> 0.99
$d_5^{\bar{1}}$ vs. $d_5^{\bar{4}}$	0.14	> 0.99	> 0.99	0.90	0.85
$d_5^{\bar{1}}$ vs. $d_5^{\bar{5}}$	0.90	> 0.99	> 0.99	> 0.99	> 0.99
$d_5^{\bar{1}}$ vs. $d_5^{\bar{6}}$	0.58	> 0.99	> 0.99	> 0.99	> 0.99
$d_5^{\bar{2}}$ vs. $d_5^{\bar{3}}$	0.36	> 0.99	> 0.99	0.99	> 0.99
$d_5^{\bar{2}}$ vs. $d_5^{\bar{4}}$	0.05	0.78	0.72	0.55	0.52
$d_5^{\bar{2}}$ vs. $d_5^{\bar{5}}$	0.70	> 0.99	> 0.99	> 0.99	> 0.99
$d_5^{\bar{2}}$ vs. $d_5^{\bar{6}}$	0.97	> 0.99	> 0.99	> 0.99	> 0.99
$d_5^{\bar{3}}$ vs. $d_5^{\bar{4}}$	0.44	> 0.99	> 0.99	0.99	> 0.99
$d_5^{\bar{3}}$ vs. $d_5^{\bar{5}}$	0.57	> 0.99	> 0.99	> 0.99	0.99
$d_5^{\bar{3}}$ vs. $d_5^{\bar{6}}$	0.35	> 0.99	> 0.99	0.99	0.99
$d_5^{\bar{4}}$ vs. $d_5^{\bar{5}}$	0.13	> 0.99	> 0.99	0.87	0.82
$d_5^{\bar{4}}$ vs. $d_5^{\bar{6}}$	0.04	0.62	0.62	0.46	0.46
$d_5^{\bar{5}}$ vs. $d_5^{\bar{6}}$	0.67	> 0.99	> 0.99	> 0.99	0.99

4.2. SIMULATION STUDY

4.2.1. Simulated G Agreement Model Assuming Marginal Homogeneity

Generation of Simulated Tables. One thousand 2^6 contingency tables were generated under the assumption of marginal homogeneity using the parameter estimates for the G model shown in Table 11, column 2. The total number of counts ranged from 32 to 116

(Table 16); the mode of 70 is similar to the observed sample size (68) of the intestinal biopsy data.

Table 16. Descriptive Statistics of Sample Size (Total Counts) of the 1000 2^6 Contingency Tables Simulated under the Assumption of Marginal Homogeneity

	Marginal Homogeneity				
Scenario	G	GP	GHeP-rog	GHeP-atyp4a	GHeP-atyp4b
Minimum	32	36	41	36	34
Maximum	116	113	125	123	115
Mode*	70	65	62, 67	67, 70	75

* Two values indicates a bi-modal distribution

One example of the simulated cell counts of the 64 possible rating patterns for the generated 2^6 contingency tables was presented in Table 12 (col. 2). The rater agreement characteristics across the 1,000 simulated contingency tables for the G agreement model are summarized in Table 17. Because the 1,000 contingency tables were generated assuming homogeneous and not category-specific global agreement, each rater's mean marginal proportion of rating presence of mucosecretion diminution should be approximately 50% (col. 2 and col. 3). Because these data were simulated from estimates based on the mucosecretion diminution data, the marginal percentage for global agreement (col 4 in Table 16) should approximate the comparable summary for the observed data (44.1%) in Table 5; under the assumed model of homogeneous global agreement, both G and GP agreement are split equally between the absence and presence

Table 17. Marginal Percentages for Different Category Specific Agreement Patterns and Rater Exclusion for the G Agreement Model Simulated under the Assumption of Marginal Homogeneity

Global Agreement Model – Marginal Homogeneity									
Rater i	Marginal % for Absence	Marginal % for Presence	G %, d_6	G on Absence %, $d_{6,0}$	G for Presence %, $d_{6,1}$	GP %, d_5	GP for Absence %, $d_{5,0}$	GP for Presence %, $d_{5,1}$	Excluded Rater, % $d_5^{\bar{i}}$
	Mean Marginal % (SD) [min,max]		**Mean Proportion (SE) [min,max]**						
1	50.0 (6.1) [32.8,74.0]	50.0 (6.1) [26.0,67.2]	44.6 (0.28) [19.1, 72.7]	22.3 (0.19) [6.3, 46.9]	22.3 (0.19) [6.3, 42.6]	10.7 (0.12) [1.1, 23.6]	5.3 (0.1) [0,17.1]	5.4 (0.1) [0,18.4]	1.8 (0.05) [0,8.8]
2	50.2 (5.9) [29.3,67.9]	49.7 (5.9) [32.1,70.7]							1.8 (0.05) [0,8.2]
3	50.0 (6.2) [28.6,69.4]	49.9 (6.2) [30.6,71.4]							1.7 (0.05) [0,8.1]
4	49.9 (6.1) [29.0,69.1]	50.1 (6.1) [30.9,71.0]							1.8 (0.05) [0,9.5]
5	49.9 (6.2) [29.1,70.3]	50.1 (6.2) [29.7,70.9]							1.8 (0.05) [0,9.6]
6	50.1 (6.2) [29.6,67.2]	49.9 (6.2) [32.8,70.4]							1.7 (0.05) [0,9.4]

* For agreement patterns see Table 4. G= Global agreement; GP= Global and partial agreement

of mucosecretion diminution (22.3%, cols. 5 and 6; 5.3% and 5.4% in cols. 8 and 9; Table 17). All $d_5^{\bar{i}}$ (*i* =1 to 6) parameters were similar (~ 1.8%, col. 10).

The percentage of the 1,000 simulated 2^6 contingency tables having none, some or all heterogeneous partial agreement parameters with sufficient statistics equal to zero is summarized in Table 18. Each set of $d_5^{\bar{i}}$ s listed in the first column of the table are disjointed. For example, 1, 2, 3 indicates that *only* heterogeneous partial agreement parameters $d_5^{\bar{1}}$, $d_5^{\bar{2}}$, and $d_5^{\bar{3}}$ had sufficient statistics equaling zero. Consequently, only pair-wise comparisons between $d_5^{\bar{4}}$, $d_5^{\bar{5}}$, and $d_5^{\bar{6}}$ are made. Regardless of how many pair-wise comparisons are made, the Type I error of each simulation is fixed at 0.05. Model G had relatively few (13.7%) simulated tables with no sufficient statistic for a heterogeneous partial agreement parameter equal to zero. This is expected, because under the G model, non-global agreement was spread uniformly across the table rather than being concentrated near the diagonal (as in GP agreement). The sufficient statistic for the heterogeneous partial agreement parameter $d_5^{\bar{4}}$ was zero in 5.7% (57) of the 1,000 simulated contingency tables. Both rating patterns representative of heterogeneous partial agreement for a particular rater must have a count of 0 for the sufficient statistic of the corresponding GHeP parameter to be zero. Because each set of 1,000 simulated 2^6 contingency tables included tables where some GHeP parameters had sufficient statistics equal to zero, the actual number of possible pair-wise comparisons was less than 1,000.

Table 18. Percent of the 1,000 Homogeneous Simulated 2^6 Contingency Tables Having None, Some or All of Its Heterogeneous Partial Agreement Parameters With Sufficient Statistic Equal to Zero

Sufficient Statistic of GHeP Parameter = 0 in Model $d_5^{\overline{i}} = 0,\ i =$	**G**	**GP**	**GHeP-rog**	**GHeP-atyp4a**	**GHeP-atyp4b**
None	13.7	69.1	56.6	36.8	54.2
6	5.2	5.0	4.4	2.8	5.0
5	4.2	2.8	4.2	6.7	5.7
4	5.7	4.0	0.5	0.5	0.1
3	4.3	4.2	6.4	19.4	6.1
2	5.8	3.6	5.5	2.8	5.1
1	3.6	4.1	5.5	8.0	5.7
1, 2	2.3	0.5	1.2	1.0	1.2
1, 3	2.0	0.6	1.8	4.7	1.3
1, 4	1.9	0.4	0.1	0.1	0
1, 5	2.2	0.4	2.4	1.7	1.4
1, 6	3.4	0.3	1.2	1.0	1.8
2, 3	1.9	0.4	0.5	2.30	1.4
2, 4	2.4	0.3	0	0	0
2, 5	1.6	0.3	0.8	0.6	0.9
2, 6	2.0	0.4	1.1	0.7	1.0
3, 4	2.4	0.1	0.1	0.1	0.1
3, 5	2.3	0.2	1.5	4.4	1.5
3, 6	2.4	0.6	1.5	1.8	1.1
4, 5	1.8	0.8	0.2	0	0
5, 6	1.8	0.4	1.2	0	0.9
1, 2, 3	0.8	0	0.4	0.7	0.6
1, 2, 4	0.6	0.1	0	0.4	0
1, 2, 5	0.6	0.1	0.6	0	0.4
1, 2, 6	0.6	0.1	0.2	0.1	0.1
2, 3, 4	0.6	0.1	0.1	0.1	0
2, 3, 5	0.6	0	0.2	0	0.3
2, 3, 6	1.4	0	0.3	0.4	0.8
3, 4, 5	0.8	0.1	0	0.3	0
3, 4, 6	0.7	0.1	0	0	0
1, 3, 4	0.7	0	0	0.1	0
1, 3, 5	0.7	0	0.1	0	0.5
1, 3, 6	1.1	0	0.2	0.8	0.4
1, 4, 5	1.0	0	0	0.5	0
1, 5, 6	1.2	0.1	0.1	0	0.6
1, 4, 6	0.9	0.1	0	0.1	0

Table 18 (continued)

Sufficient Statistic of GHeP Parameter = 0 in Model $d_5^{\bar{i}} = 0,\ i =$	**G**	**GP**	**GHeP-rog**	**GHeP-atyp4a**	**GHeP-atyp4b**
2, 4, 5	0.9	0.1	0	0	0
2, 4, 6	1.2	0.1	0	0.1	0.1
4, 5, 6	0.9	0	0	0	0
3, 5, 6	1.1	0	0.1	0	0.2
2, 5, 6	0.5	0	0.4	0.2	0.5
3, 4, 5, 6	0.3	0	0	0.1	0
2, 4, 5, 6	0.6	0	0	0	0
2, 3, 5, 6	0.4	0	0.1	0	0.1
2, 3, 4, 6	0.8	0.1	0	0.1	0
2, 3, 4, 6	0.2	0	0	0	0
1, 4, 5, 6	0.5	0	0	0	0
1, 3, 5, 6	0.6	0	0	0	0.3
1, 3, 4, 6	0.5	0	0.1	0.1	0
1, 3, 4, 5	0.4	0	0	0	0
1, 2, 5, 6	0.4	0	0	0.1	0.1
1, 2, 4, 6	0.2	0	0	0	0
1, 2, 4, 6	0.9	0	0	0	0
1, 2, 3, 5	0.2	0	0.2	0	0.3
1, 2, 3, 4	0.2	0	0.2	0	0.1
2, 3, 4, 5, 6	0.6	0	0	0.4	0
1, 3, 4, 5, 6	0.2	0	0	0	0
1, 2, 4, 5, 6	0.1	0	0	0	0
1, 3, 4, 5, 6	0.2	0.1	0	0	0
1, 2, 3, 5, 6	0.4	0	0	0	0.1
1, 2, 3, 4, 6	0.2	0	0	0	0
1, 2, 3, 4, 5	0.6	0	0	0	0
All	0.2	0	0	0	0
TOTAL	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**

Analysis Assuming Marginal Homogeneity. The number of times each possible pair-wise comparison was statistically significant across the 1,000 simulated contingency tables is presented in Table 19. Shaded comparisons highlight the pair-wise comparisons involving $d_5^{\bar{4}}$. None of the fifteen heterogeneous partial agreement parameter pair-wise comparisons was significant in any table generated under the G agreement model, either unadjusted or adjusted for multiple comparisons.

Table 19. Number of Times Each Possible Pair-wise Comparison was Statistically Significant Across 1000 Tables Simulated under the G Agreement Model with Marginal Homogeneity

Comparison	Unadjusted	Bonferroni	Holm's - Bonferroni	Sidak	Holm's - Sidak
	n (%)				
$\overline{\hat{d}_5^1}$ vs. $\overline{\hat{d}_5^2}$	0	0	0	0	0
$\overline{\hat{d}_5^1}$ vs. $\overline{\hat{d}_5^3}$	0	0	0	0	0
$\overline{\hat{d}_5^1}$ vs. $\overline{\hat{d}_5^4}$	0	0	0	0	0
$\overline{\hat{d}_5^1}$ vs. $\overline{\hat{d}_5^5}$	0	0	0	0	0
$\overline{\hat{d}_5^1}$ vs. $\overline{\hat{d}_5^6}$	0	0	0	0	0
$\overline{\hat{d}_5^2}$ vs. $\overline{\hat{d}_5^3}$	0	0	0	0	0
$\overline{\hat{d}_5^2}$ vs. $\overline{\hat{d}_5^4}$	0	0	0	0	0
$\overline{\hat{d}_5^2}$ vs. $\overline{\hat{d}_5^5}$	0	0	0	0	0
$\overline{\hat{d}_5^2}$ vs. $\overline{\hat{d}_5^6}$	0	0	0	0	0
$\overline{\hat{d}_5^3}$ vs. $\overline{\hat{d}_5^4}$	0	0	0	0	0
$\overline{\hat{d}_5^3}$ vs. $\overline{\hat{d}_5^5}$	0	0	0	0	0
$\overline{\hat{d}_5^3}$ vs. $\overline{\hat{d}_5^6}$	0	0	0	0	0
$\overline{\hat{d}_5^4}$ vs. $\overline{\hat{d}_5^5}$	0	0	0	0	0
$\overline{\hat{d}_5^4}$ vs. $\overline{\hat{d}_5^6}$	0	0	0	0	0
$\overline{\hat{d}_5^5}$ vs. $\overline{\hat{d}_5^6}$	0	0	0	0	0

4.2.2. Simulated GP Agreement Model Assuming Marginal Homogeneity

Generation of Simulated Tables. One thousand 2^6 contingency tables were generated for the GP model under the assumption of marginal homogeneity using the parameter estimates shown in Table 11, column 3. The total number of counts per table ranged from 36 to 113 with a mode of 65 (Table 16). An example of the simulated cell counts for a generated 2^6 contingency tables for this GP scenario was presented in Table 12 (col.3).

A summary of the rater agreement characteristics across the 1,000 simulated contingency tables for the GP agreement model is presented in Table 20. As in the G model, global agreement in the GP model is not category specific, so the marginal distributions for presence and absence and the global agreement estimates in Table 20 are similar to those for the G model in Table 17. However, relatively more observations (25.6%) represent partial agreement for the GP model; this partial agreement is split equally between agreement on presence and absence of mucosecretion diminution. The percentage of five-way agreement is similar when each rater is excluded (~ 4.2%, col. 10).

The simulation under the GP agreement model had the highest percent (69.1%) of contingency tables among the models considered with no sufficient statistics for heterogeneous partial agreement parameters equal to zero (Table 18). This is expected because this model concentrates the counts on the main diagonal (rating patterns (000000) and (111111)) and equally across the ten rating patterns representing five-way agreement on the immediate off-diagonal.

Table 20. Marginal Percentages for Different Category Specific Agreement Patterns and Rater Exclusion for the GP Agreement Model Simulated under the Assumption of Marginal Homogeneity

Global & Partial Agreement Model – Marginal Homogeneity									
Rater *i*	Marginal % for Absence	Marginal % for Presence	G %, d_6	G for Absence %, $d_{6,0}$	G for Presence %, $d_{6,1}$	GP %, d_5	GP for Absence %, $d_{5,0}$	GP for Presence %, $d_{5,1}$	Excluded Rater, % $d_5^{\bar{i}}$
	Mean Marginal % (SE) [min,max]		**Mean Proportion (SE) [min,max]**						
1	50.1 (6.3) [25.0,66.7]	49.8 (6.3) [33.3,75.0]	43.5 (0.27) [19.0,69.7]	21.6 (0.19) [3.9,43.1]	21.9 (0.18) [4.7,44.4]	25.6 (0.23) [1.5,52.0]	12.9 (0.16) [0,31.1]	12.6 (0.15) [0,29.6]	4.2 (0.1) [0,14.0]
2	50.0 (6.0) [33.3,67.1]	50.0 (6.0) [32.9,66.7]							4.3 (0.8) [0,16.2]
3	49.8 (5.9) [33.3,67.7]	50.2 (5.9) [32.3,66.7]							4.3 (0.1) [0,15.0]
4	49.8 (6.2) [25.0,68.6]	50.2 (6.2) [31.4,75.0]							4.2 (0.1) [0,16.3]
5	49.6 (6.0) [25.0,72.3]	50.3 (6.0) [27.7,75.0]							4.3 (0.1) [0,14.6]
6	49.9 (6.1) [29.9,67.6]	50.1 (6.1) [32.4,70.1]							4.2 (0.1) [0,13.1]

* For agreement patterns see Table 4. G= Global agreement; GP= Global and partial agreement

Analysis Assuming Marginal Homogeneity. The number of times each possible pair-wise comparison was statistically significant across the 1,000 simulated contingency tables is presented in Table 21. Shaded rows highlight the pair-wise comparisons involving $d_{5}^{\bar{4}}$. None of the fifteen heterogeneous partial agreement parameter pair-wise comparisons was significant in any table generated under the GP agreement model, either unadjusted or adjusted for multiple comparisons.

Table 21. Number of Times Each Possible Pair-wise Comparison was Statistically Significant Across 1000 Tables Simulated under the GP Agreement Model with Marginal Homogeneity

Comparison	Unadjusted	Bonferroni	Holm's - Bonferroni	Sidak	Holm's - Sidak
	n (%)				
$\hat{d}_{5}^{\bar{1}}$ vs. $\hat{d}_{5}^{\bar{2}}$	0	0	0	0	0
$\hat{d}_{5}^{\bar{1}}$ vs. $\hat{d}_{5}^{\bar{3}}$	0	0	0	0	0
$\hat{d}_{5}^{\bar{1}}$ vs. $\hat{d}_{5}^{\bar{4}}$	0	0	0	0	0
$\hat{d}_{5}^{\bar{1}}$ vs. $\hat{d}_{5}^{\bar{5}}$	0	0	0	0	0
$\hat{d}_{5}^{\bar{1}}$ vs. $\hat{d}_{5}^{\bar{6}}$	0	0	0	0	0
$\hat{d}_{5}^{\bar{2}}$ vs. $\hat{d}_{5}^{\bar{3}}$	0	0	0	0	0
$\hat{d}_{5}^{\bar{2}}$ vs. $\hat{d}_{5}^{\bar{4}}$	0	0	0	0	0
$\hat{d}_{5}^{\bar{2}}$ vs. $\hat{d}_{5}^{\bar{5}}$	0	0	0	0	0
$\hat{d}_{5}^{\bar{2}}$ vs. $\hat{d}_{5}^{\bar{6}}$	0	0	0	0	0
$\hat{d}_{5}^{\bar{3}}$ vs. $\hat{d}_{5}^{\bar{4}}$	0	0	0	0	0
$\hat{d}_{5}^{\bar{3}}$ vs. $\hat{d}_{5}^{\bar{5}}$	0	0	0	0	0
$\hat{d}_{5}^{\bar{3}}$ vs. $\hat{d}_{5}^{\bar{6}}$	0	0	0	0	0
$\hat{d}_{5}^{\bar{4}}$ vs. $\hat{d}_{5}^{\bar{5}}$	0	0	0	0	0
$\hat{d}_{5}^{\bar{4}}$ vs. $\hat{d}_{5}^{\bar{6}}$	0	0	0	0	0
$\hat{d}_{5}^{\bar{5}}$ vs. $\hat{d}_{5}^{\bar{6}}$	0	0	0	0	0

4.2.3. Simulated GHeP-rog Agreement Model Assuming Marginal Homogeneity

Generation of Simulated Tables. One thousand 2^6 contingency tables were generated under the assumption of marginal homogeneity using the parameter estimates for the GHeP-rog model shown in Table 11, column 4. The total number of counts per table ranged from 41 to 125 with modes of 62 and 67 counts (Table 16). An example of the simulated cell counts for a generated 2^6 contingency tables for this GHeP-rog scenario was presented in Table 12 (col.4).

The rater agreement characteristics across the 1,000 simulated contingency tables for this GHeP-rog agreement model are presented in Table 22. The global agreement and partial agreement from the 1,000 simulated contingency tables are comparable to those observed from the intestinal biopsy data (Table 5). The global agreement of the simulated data was 42.2% vs. 44.1% in the observed data, and partial agreement was 27.9% vs. 25.0%. Because the data were simulated under the assumption of non-category specific global or partial agreement, the marginal percentages for the absence and presence of mucosecretion diminution should be similar for global and partial agreement (global agreement, absence: 21.4% , presence: 20.8%; partial agreement, absence: 13.9%, presence: 14.0%).

The marginal percentage of five-way agreement when a specific rater is excluded from the simulated data differed slightly from that of the observed data because of the assumption of marginal homogeneity. The mean marginal percentage of five-way agreement when Rater 1, 2, 3, 4, 5, or 6 is excluded from the simulated data is 3.5%, 4.8%, 2.4%, 9.0%, 3.5%, and 4.9%, respectively, vs. 2.9%, 4.4%, 1.5%, 8.8%, 2.9%, and 4.4%, respectively, observed from the mucosecretion diminution data. Five-way agreement was highest when Rater 4 was excluded.

Table 22. Marginal Percentages for Different Category Specific Agreement Patterns and Rater Exclusion for the GHeP-rog Agreement Model Simulated under the Assumption of Marginal Homogeneity

GHeP-rog Model – Marginal Homogeneity									
Rater *i*	Marginal % for Absence	Marginal % for Presence	G %, d_6	G for Absence %, $d_{6,0}$	G for Presence %, $d_{6,1}$	GP %, d_5	GP for Absence %, $d_{5,0}$	GP for Presence %, $d_{5,1}$	Excluded Rater, % $d_5^{\bar{i}}$
	Mean Marginal % (SE) [min,max]		**Mean Proportion (SE) [min,max]**						
1	50.0 (6.2) [29.6,68.5]	49.9 (6.2) [31.5,70.4]	42.2 (0.26) [17.1,70.0]	21.4 (0.18) [6.3,44.6]	20.8 (0.18) [4.8,40.3]	27.9 (0.25) [6.0,60.0]	13.9 (0.16) [0,33.3]	14.0 (0.16) [1.4,41.4]	3.5 (0.1) [0,20.0]
2	50.3 (5.7) [30.7,69.4]	49.7 (5.7) [30.6,69.3]							4.8 (0.1) [0,25.6]
3	50.2 (5.9) [31.5,70.8]	49.8 (5.8) [29.2,68.5]							2.4 (0.1) [0,25.8]
4	50.2 (5.8) [31.1,69.8]	49.7 (5.8) [30.2,68.9]							9.0 (0.2) [0,36.1]
5	50.0 (5.9) [28.4,70.7]	49.9 (5.9) [29.3,71.6]							3.5 (0.1) [0,21.4]
6	50.4 (5.9) [25.7,70.8]	49.6 (5.9) [29.2,74.3]							4.9 (0.1) [0,25.7

For agreement patterns see Table 4. G= Global agreement; GP= Global and partial agreement

The sufficient statistic for the heterogeneous partial agreement parameter $d_5^{\bar{4}}$ was zero in only 0.5% (5) of these 1,000 simulated contingency tables (Table 18, row 4, col. 4), and 44.4% of the contingency tables had at least one heterogeneous partial agreement parameter with a sufficient statistic equal to zero.

Analysis Assuming Marginal Homogeneity. The number of times each possible pair-wise comparison was statistically significant across the 1,000 simulated contingency tables is presented in Table 23. Results are presented by subsets of simulations defined by the number of possible pair-wise comparisons (15, 10, 6, 3, or 1). For subsets of size less than 15, “Missing” denotes the number of simulations in which the given comparison was not possible due to sampling zeros.

The vast majority of significant unadjusted pair-wise comparisons involved Rater 4, indicating that five-way agreement when Rater 4 is excluded is different from five-way agreement when the other raters are excluded. For pair-wise comparisons involving Rater 4, the number of adjusted significant pair-wise comparisons was reduced to 2 or less. The few statistically significant adjusted pair-wise comparisons all involved Rater 4. In this simulation scenario, Rater 4 is designated as the atypical rater.

Table 23. Number (%) of Times Each Possible Pair-wise Comparison was Statistically Significant Across 1000 Tables Simulated under the GHeP-rog Agreement Model with Marginal Homogeneity

Number of Pair-wise Comparisons (N)	**15 (556)**	**10 (265)**	**6 (136)**	**3 (27)**	**1 (6)**
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{2}}$					
Unadjusted	3 (0.5)	2 (1.3)	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	110	91	26	6
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{3}}$					
Unadjusted	3 (0.5)	1 (0.6)	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	119	103	23	6
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{4}}$					
Unadjusted	80 (14.1)	38 (18.5)	16 (24.2)	2 (20.0)	0
Bonferroni	1 (0.2)	0	1 (1.5)	0	0
Holm's- Bonferroni	1 (0.2)	0	1 (1.5)	0	0
Sidak	1 (0.2)	0	1 (1.5)	0	0
Holm's-Sidak	1 (0.2)	0	1 (1.5)	0	0
Missing	0	60	70	17	5
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{5}}$					
Unadjusted	4 (0.7)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	97	104	23	6

Table 23 (continued)

Possible Pair-wise Comparisons (N)	**15 (556)**	**10 (265)**	**6 (136)**	**3 (27)**	**1 (6)**
$\hat{d}_5^{\overline{1}}$ vs. $\hat{d}_5^{\overline{6}}$					
Unadjusted	5 (0.8)	2 (1.2)	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	99	105	24	6
$\hat{d}_5^{\overline{2}}$ vs. $\hat{d}_5^{\overline{3}}$					
Unadjusted	4 (0.7)	2 (1.4)	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	119	85	26	6
$\hat{d}_5^{\overline{2}}$ vs. $\hat{d}_5^{\overline{4}}$					
Unadjusted	81 (14.3)	32 (15.6)	24 (25.0)	0	--
Bonferroni	1 (0.2)	1 (0.5)	1 (1.0)	0	--
Holm's- Bonferroni	1 (0.2)	1 (0.5)	2 (2.0)	0	--
Sidak	1 (0.2)	1 (0.5)	1 (1.0)	0	--
Holm's-Sidak	1 (0.2)	1 (0.5)	2 (2.0)	0	--
Missing	0	60	40	22	6
$\hat{d}_5^{\overline{2}}$ vs. $\hat{d}_5^{\overline{5}}$					
Unadjusted	3 (0.5)	0	0	0	0
Bonferroni	0	0	0	0	0
Holm's- Bonferroni	0	0	0	0	0
Sidak	0	0	0	0	0
Holm's-Sidak	0	0	0	0	0
Missing	0	97	89	25	5

Table 23 (continued)

Possible Pair-wise Comparisons (N)	**15 (556)**	**10 (265)**	**6 (136)**	**3 (27)**	**1 (6)**
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{6}}$					
Unadjusted	4 (0.7)	1 (0.6)	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	99	75	26	6
$\hat{d}_5^{\bar{3}}$ vs. $\hat{d}_5^{\bar{4}}$					
Unadjusted	85 (15.0)	28 (14.3)	21 (26.5)	1 (7.7)	--
Bonferroni	2 (0.4)	1 (0.5)	0	0	--
Holm's- Bonferroni	2 (0.4)	1 (0.5)	0	0	--
Sidak	2 (0.4)	1 (0.5)	0	0	--
Holm's-Sidak	2 (0.4)	1 (0.5)	0	0	--
Missing	0	69	57	14	6
$\hat{d}_5^{\bar{3}}$ vs. $\hat{d}_5^{\bar{5}}$					
Unadjusted	5 (0.8)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	106	100	25	6
$\hat{d}_5^{\bar{3}}$ vs. $\hat{d}_5^{\bar{6}}$					
Unadjusted	0	0	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	108	89	21	6

Table 23 (continued)

Possible Pair-wise Comparisons (N)	**15** **(556)**	**10** **(265)**	**6** **(136)**	**3** **(27)**	**1** **(6)**
$\hat{d}_{5}^{\overline{4}}$ vs. $\hat{d}_{5}^{\overline{5}}$					
Unadjusted	78 (13.7)	27 (12.4)	15 (20.6)	3 (27.3)	0
Bonferroni	1 (0.2)	1 (0.4)	1 (1.4)	0	0
Holm's- Bonferroni	1 (0.2)	1 (0.4)	1 (1.4)	0	0
Sidak	1 (0.2)	1 (0.4)	1 (1.4)	0	0
Holm's-Sidak	1 (0.2)	1 (0.4)	1 (1.4)	0	0
Missing	0	47	63	16	4
$\hat{d}_{5}^{\overline{4}}$ vs. $\hat{d}_{5}^{\overline{6}}$					
Unadjusted	89 (15.7)	38 (17.6)	24 (29.3)	4 (30.8)	1 (50.0)
Bonferroni	1 (0.2)	1 (0.4)	2 (2.4)	0	0
Holm's- Bonferroni	1 (0.2)	1 (0.4)	2 (2.4)	0	0
Sidak	1 (0.2)	1 (0.4)	2 (2.4)	0	0
Holm's-Sidak	1 (0.2)	1 (0.4)	2 (2.4)	0	0
Missing	0	49	54	14	4
$\hat{d}_{5}^{\overline{5}}$ vs. $\hat{d}_{5}^{\overline{6}}$					
Unadjusted	2 (0.4)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	86	99	22	6

For each simulated 2^6 contingency table that had at least one significant p-value, the proportion of comparisons involving Rater 4 was assessed (Table 24). Table 24 summarizes the power to detect Rater 4 as atypical when "atypical" was defined as being different from one, two, …, five raters or from at least one other rater. The power to identify Rater 4 as being atypical is 27.7% using a criterion of at least one of the five unadjusted pair-wise comparisons involving $\hat{d}_5^{\bar{4}}$ is statistically significant. The power is reduced to 0.6% when the analysis is adjusted for the number of comparisons. Power is similarly low across the four multiple comparison procedures considered using the criterion that at least one of the five pair-wise comparisons involving $\hat{d}_5^{\bar{4}}$ is statistically significant.

Table 24. Power to Identify Rater 4 as the Atypical Rater Using Various Criteria by Multiple Comparison Procedure for the GHeP-rog Scenario

	Multiple Comparison Procedure				
Rater 4 Differs from:	**Unadjusted**	**Bonferroni**	**Holm's Bonferroni**	**Sidak**	**Holm's Sidak**
One rater	0.078	0.002	0.002	0.002	0.002
Two raters	0.070	0.001	0	0.001	0
Three raters	0.066	0.001	0.002	0.001	0.002
Four raters	0.040	0.001	0.001	0.001	0.001
Five raters	0.023	0.001	0.001	0.001	0.001
At least one rater	0.277	0.006	0.006	0.006	0.006

The proportion of simulations that identified a rater other than Rater 4 as the atypical rater was 0.007 based upon unadjusted p-values and 0 when based on adjusted p-values (Table 25). The overall probability that *any* rater is identified as atypical is approximately 6% based on unadjusted comparisons and less than 1% if adjustments are made (Table 26). However, the probability that Rater 4 is identified given that an atypical rater is identified is greater than 93% (Table 27).

Table 25. Proportion of Simulations that Identify a Rater Other Than Rater 4 as the Atypical Rater by Multiple Comparison Procedure for the Three Scenarios Simulated under the Assumption of Marginal Homogeneity

	Analytic Approach: Marginal Homogeneity				
Simulation Scenario	Unadjusted	Bonferroni	Holm's-Bonferroni	Sidak	Holm's-Sidak
GHeP-rog	0.007	0	0	0	0
GHeP-atyp4a	0.171	0.019	0.019	0.019	0.019
GHeP-atyp4b	0.026	0	0	0	0

Table 26. Overall Probability (%) of Identifying any Rater as the Atypical Rater for Data Simulated Under the Assumption of Marginal Homogeneity

Model	**Unadjusted**	**Bonferroni**	**Holm's-Bonferroni**	**Sidak**	**Holm's-Sidak**
GHeP-rog	6.05	0.13	0.14	0.13	0.14
GHeP-atyp4a	9.14	0.86	0.94	0.86	0.94
GHeP-atyp4b	10.56	0.63	0.65	0.63	0.70

Table 27. Conditional Probability (%) of Identifying the Designated Atypical Rater as Atypical for Data Simulated Under the Assumption of Marginal Homogeneity

Model	**Unadjusted**	**Bonferroni**	**Holm's-Bonferroni**	**Sidak**	**Holm's-Sidak**
GHeP-rog	94.37	100	100	100	100
GHeP-atyp4a	60.65	53.76	55.45	53.76	55.45
GHeP-atyp4b	97.11	100	100	100	100

4.2.4. Simulated GHeP-atyp4a Agreement Model Assuming Marginal Homogeneity

Generation of Simulated Tables. One thousand 2^6 contingency tables were generated under the assumption of marginal homogeneity using the parameter estimates for the GHeP-atyp4a model shown in Table 11, column 5. The total number of counts per table ranged from 36 to 123 with modes of 67 and 70 counts (Table 16). An example of the simulated cell counts for the generated 2^6 contingency tables for this GHeP-atyp4a scenario was presented in Table 12 (col. 5).

A summary of the rater agreement characteristics across the 1,000 simulated contingency tables for the GHeP-atyp4a agreement model is presented in Table 28. The percentages of five-way agreement when Raters 1, 2, 3, 5, and 6 are excluded should be similar and less than the percentage of five-way agreement when Rater 4 is excluded. The percentages of five-way agreement when Raters 1, 2, 3, 5, and 6 are excluded are ~3.5%, and the percentage of five-way agreement when Rater 4 is excluded is 9.0%.

Table 28. Marginal Percentages for Different Category Specific Agreement Patterns and Rater Exclusion for the GHeP-atyp4a Model Simulated under the Assumption of Marginal Homogeneity

GHeP-atyp4a Model– Marginal Homogeneity									
Rater i	Marginal % for Absence	Marginal % for Presence	G %, d_6	G for Absence %, $d_{6,0}$	G for Presence %, $d_{6,1}$	GP %, d_5	GP for Absence %, $d_{5,0}$	GP for Presence %, $d_{5,1}$	Excluded Rater, % $d_5^{\bar{i}}$
	Mean Marginal % (SE) [min,max]		**Mean Proportion (SE) [min,max]**						
1	50.2 (6.1) [27.5,67.1]	49.8 (6.1) [32.9,72.5]	43.1 (0.26) [16.9,70.2]	21.7 (0.18) [4.1,43.2]	21.4 (0.18) [5.6,40.7]	26.2 (0.25) [6.9,54.5]	13.2 (0.16) [0,30.7]	13.0 (0.16) [1.4,30.2]	3.4 (0.1) [0,16.9]
2	50.2 (6.0) [31.0,72.5]	49.7 (6.0) [27.5,69.0]							3.5 (0.1) [0,16.3]
3	50.3 (5.9) [31.2,73.4]	49.6 (5.9) [26.6,68.8]							3.4 (0.1) [0,13.1]
4	50.2 (6.0) [28.2,70.9]	49.8 (6.0) [29.1,71.7]							9.0 (0.1) [0,26.2]
5	50.1 (6.2) [29.2,67.5]	49.8 (6.2) [32.5,70.8]							3.4 (0.1) [0,14.6]
6	50.2 (6.0) [33.3,72.4]	49.8 (6.0) [27.6,66.7]							3.5 (0.1) [0,15.0]

* For agreement patterns see Table 4. G= Global agreement; GP= Global and partial agreement

In the GHeP-atyp4a simulated scenario, only 5 of the 1, 000 simulated 2^6 contingency tables had counts equal to zero for rating patterns (000100) and (111011), i.e., $\hat{d}_5^{\bar{4}} = 0$, and 63.2% of the contingency tables had at least one heterogeneous partial agreement parameter with a sufficient statistic equal to zero.

Analysis Assuming Marginal Homogeneity. The number of times each possible pair-wise comparison was statistically significant across the 1,000 simulated contingency tables is shown in Table 29. Within each column, relatively more statistically significant unadjusted pair-wise comparisons involved Rater 4. Very few of the adjusted pair-wise comparisons were statistically significant. In contrast to the GHeP-rog simulation scenario, a sizeable number of statistically significant unadjusted pair-wise comparisons did not involve $d_5^{\bar{4}}$. This explains why the unadjusted conditional probability of identifying the designated atypical rater as atypical for the GHeP-atyp4a (60.65%) scenario is less than that from the GHeP-rog scenario (94.37%, Table 27).

Table 29. Number (%) of Times Each Possible Pair-wise Comparison was Statistically Significant Across 1000 Tables Simulated under the GHeP-atyp4a Agreement Model with Marginal Homogeneity

Possible Pair-wise Comparisons	**15**	**10**	**6**	**3**	**1**
(N)	**(368)**	**(402)**	**(191)**	**(32)**	**(7)**
$\hat{d}_{5}^{\overline{1}}$ vs. $\hat{d}_{5}^{\overline{2}}$					
Unadjusted	26 (7.1)	12 (4.1)	7 (10.0)	0	--
Bonferroni	0	1 (0.3)	2 (2.8)	0	--
Holm's- Bonferroni	0	1 (0.3)	2 (2.8)	0	--
Sidak	0	1 (0.3)	2 (2.8)	0	--
Holm's-Sidak	0	1 (0.3)	2 (2.8)	0	--
Missing	0	108	121	30	7
$\hat{d}_{5}^{\overline{1}}$ vs. $\hat{d}_{5}^{\overline{3}}$					
Unadjusted	21(5.7)	6 (4.7)	3 (15.0)	0	--
Bonferroni	2 (0.5)	2 (1.6)	0	0	--
Holm's- Bonferroni	2 (0.5)	2 (1.6)	0	0	--
Sidak	2 (0.5)	2 (1.6)	0	0	--
Holm's-Sidak	2 (0.5)	2 (1.6)	0	0	--
Missing	0	274	171	30	7
$\hat{d}_{5}^{\overline{1}}$ vs. $\hat{d}_{5}^{\overline{4}}$					
Unadjusted	79(21.5)	24 (7.6)	8 (17.1)	0	1(100)
Bonferroni	5 (1.4)	1 (0.3)	2 (1.9)	0	1(100)
Holm's- Bonferroni	6 (1.6)	1 (0.3)	2 (1.9)	0	1(100)
Sidak	5 (1.4)	1 (0.3)	2 (1.9)	0	1(100)
Holm's-Sidak	6 (1.6)	1 (0.3)	2 (1.9)	0	1(100)
Missing	0	85	86	22	6
$\hat{d}_{5}^{\overline{1}}$ vs. $\hat{d}_{5}^{\overline{5}}$					
Unadjusted	24(6.5)	5 (2.0)	3 (6.1)	0	--
Bonferroni	0	1 (0.2)	0	0	--
Holm's- Bonferroni	0	1 (0.2)	0	0	--
Sidak	0	1 (0.2)	0	0	--
Holm's-Sidak	0	1 (0.2)	0	0	--
Missing	0	147	142	29	7

Table 29 (continued)

Possible Pair-wise Comparisons (N)	**15 (368)**	**10 (402)**	**6 (191)**	**3 (32)**	**1 (7)**
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{6}}$					
Unadjusted	28 (7.6)	10 (3.4)	3 (4.1)	0	--
Bonferroni	3 (0.8)	3 (1.0)	1 (1.4)	0	--
Holm's- Bonferroni	3 (0.8)	3 (1.0)	2 (2.7)	0	--
Sidak	3 (0.8)	3 (1.0)	1 (1.4)	0	--
Holm's-Sidak	3 (0.8)	3 (1.0)	2 (2.7)	0	--
Missing	0	108	117	26	7
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{3}}$					
Unadjusted	23 (6.3)	14 (7.8)	2 (5.9)	0	--
Bonferroni	3 (0.8)	1 (0.5)	0	0	--
Holm's- Bonferroni	3 (0.8)	1 (0.5)	0	0	--
Sidak	3 (0.8)	1 (0.5)	0	0	--
Holm's-Sidak	3 (0.8)	1 (0.5)	0	0	--
Missing	0	222	157	31	7
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{4}}$					
Unadjusted	61 (16.5)	28 (7.6)	24 (16.7)	2(12.5)	1(100)
Bonferroni	5 (1.3)	2 (0.5)	3 (2.1)	0	0
Holm's- Bonferroni	5 (1.3)	2 (0.5)	3 (2.1)	0	0
Sidak	5 (1.3)	2 (0.5)	3 (2.1)	0	0
Holm's-Sidak	5 (1.3)	2 (0.5)	3 (2.1)	0	0
Missing	0	33	47	16	6
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{5}}$					
Unadjusted	17 (4.6)	5 (1.6)	9 (11.7)	0	--
Bonferroni	0	1 (0.3)	3 (3.9)	0	--
Holm's- Bonferroni	0	1 (0.3)	3 (3.9)	0	--
Sidak	0	1 (0.3)	3 (3.9)	0	--
Holm's-Sidak	0	1 (0.3)	3 (3.9)	0	--
Missing	0	95	114	27	7

Table 29 (continued)

Possible Pair-wise Comparisons (N)	**15 (368)**	**10 (402)**	**6 (191)**	**3 (32)**	**1 (7)**
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{6}}$					
Unadjusted	24 (6.5)	8 (2.3)	8 (7.3)	1(11.1)	0
Bonferroni	2 (0.5)	3 (0.9)	2 (1.8)	0	0
Holm's- Bonferroni	2 (0.5)	3 (0.9)	2 (1.8)	0	0
Sidak	2 (0.5)	3 (0.9)	2 (1.8)	0	0
Holm's-Sidak	2 (0.5)	3 (0.9)	2 (1.8)	0	0
Missing	0	56	81	23	6
$\hat{d}_5^{\bar{3}}$ vs. $\hat{d}_5^{\bar{4}}$					
Unadjusted	88 (23.9)	45 (22.2)	12 (21.1)	0	--
Bonferroni	4 (1.1)	3 (1.4)	1 (1.8)	0	--
Holm's- Bonferroni	6 (1.6)	3 (1.4)	2 (3.5)	0	--
Sidak	4 (1.1)	3 (1.4)	1 (1.8)	0	--
Holm's-Sidak	6 (1.6)	3 (1.4)	2 (3.5)	0	--
Missing	0	199	134	28	7
$\hat{d}_5^{\bar{3}}$ vs. $\hat{d}_5^{\bar{5}}$					
Unadjusted	24 (6.5)	9 (6.3)	1 (3.7)	1(100)	--
Bonferroni	3 (0.8)	1 (0.7)	0	0	--
Holm's- Bonferroni	3 (0.8)	1 (0.7)	0	0	--
Sidak	3 (0.8)	1 (0.7)	0	0	--
Holm's-Sidak	3 (0.8)	1 (0.7)	0	0	--
Missing	0	261	164	31	7
$\hat{d}_5^{\bar{3}}$ vs. $\hat{d}_5^{\bar{6}}$					
Unadjusted	33 (9.0)	14(7.8)	1 (3.0)	0	--
Bonferroni	4 (1.1)	0	0	0	--
Holm's- Bonferroni	4 (1.1)	0	0	0	--
Sidak	4 (1.1)	0	0	0	--
Holm's-Sidak	4 (1.1)	0	0	0	--
Missing	0	222	158	30	7

Table 29 (continued)

Possible Pair-wise Comparisons (N)	**15 (368)**	**10 (402)**	**6 (191)**	**3 (32)**	**1 (7)**
$\hat{d}_{5}^{\overline{4}}$ vs. $\hat{d}_{5}^{\overline{5}}$					
Unadjusted	68 (18.4)	23 7.0)	20 (17.2)	1(7.7)	--
Bonferroni	7 (1.9)	2 (0.6)	2 (1.7)	0	--
Holm's- Bonferroni	7 (1.9)	2 (0.6)	2 (1.7)	1(7.7)	--
Sidak	7 (1.9)	2 (0.6)	2 (1.7)	0	--
Holm's-Sidak	7 (1.9)	2 (0.6)	2 (1.7)	1(7.7)	--
Missing	0	72	75	19	7
$\hat{d}_{5}^{\overline{4}}$ vs. $\hat{d}_{5}^{\overline{6}}$					
Unadjusted	57 (15.5)	24(6.5)	22 (14.8)	1(5.9)	0
Bonferroni	6 (1.6)	3 (0.8)	3 (2.0)	0	0
Holm's- Bonferroni	7 (1.9)	3 (0.8)	3 (2.0)	0	0
Sidak	6 (1.6)	3 (0.8)	3 (2.0)	0	0
Holm's-Sidak	7 (1.9)	3 (0.8)	3 (2.0)	0	0
Missing	0	33	43	15	3
$\hat{d}_{5}^{\overline{5}}$ vs. $\hat{d}_{5}^{\overline{6}}$					
Unadjusted	34 (9.2)	7 (2.2)	4 (4.9)	1 (25.0)	--
Bonferroni	2 (0.5)	2 (0.7)	0	1 (25.0)	--
Holm's- Bonferroni	2 (0.5)	2 (0.7)	1 (1.2)	1 (25.0)	--
Sidak	2 (0.5)	2 (0.7)	0	1 (25.0)	--
Holm's-Sidak	2 (0.5)	2 (0.7)	1 (1.2)	1 (25.0)	--
Missing	0	95	109	28	7

Table 30 summarizes the power to identify Rater 4 as the atypical rater when $d_5^{\bar{4}}$ differs from one or more than one other $d_5^{\bar{i}}$. The power of the unadjusted comparisons to detect Rater 4 as different from exactly one other rater is slightly higher (11.3%) in Table 30 compared to the comparable power in Table 24 (7.8%); the overall power (27.5% vs. 27.7%, respectively) is similar. Regardless of the multiple comparison procedure used, the power to identify Rater 4 as atypical using a criterion that $d_5^{\bar{4}}$ differs from at least one $d_5^{\bar{i}}$ (i = 1, 2, 3, 5, or 6) is at most 2.3%. The increase in power from 0.6% in the GHeP-rog scenario to 2.2% in Table 30 may be explained by Raters 1, 2, 3, 5, and 6 being more homogeneous with respect to their rating characteristics in the GHeP-atyp4a scenario than in the GHeP-rog scenario.

Table 30. Power to Identify Rater 4 as the Atypical Rater Using Various Criteria By Multiple Comparison Procedure for the GHeP-atyp4a Scenario

Rater 4 Differs From:	Unadjusted	Bonferroni	Holm's-Bonferroni	Sidak	Holm's-Sidak
One rater	0.113	0.011	0.011	0.011	0.011
Two raters	0.069	0.002	0.002	0.002	0.002
Three raters	0.039	0.002	0.002	0.002	0.002
Four raters	0.039	0.006	0.005	0.006	0.005
Five raters	0.015	0.001	0.003	0.001	0.003
At least one rater	0.275	0.022	0.023	0.022	0.023

Because there were adjusted statistically significant pair-wise comparisons that did not involve $d_5^{\bar{4}}$, the proportion of simulations that identified a rater other than Rater 4 as the atypical rater was calculated. Using an identification criterion that at least one of the pair-wise comparisons between the remaining five raters had to be significant to identify a rater as being atypical, 17.1% of these simulations identified a rater other than Rater 4 as the atypical rater based upon unadjusted comparisons compared to only1.9% from adjusted comparisons (Table 25).

The overall probability that *any* rater is identified as atypical is 9.14% if adjustments for the number of comparisons are not made and less than 1% if adjustments are made (Table 26). The probability that Rater 4 is identified given that an atypical rater is identified is 60.65% based on adjusted comparisons and 54%-55% for the four multiple comparison procedures (Table 27). It is relatively more difficult to correctly identify the designated atypical rater under this scenario. A greater proportion (63.2%) of the GHeP-atyp4a simulations had at least one heterogeneous partial agreement parameter with a sufficient statistic equal to zero compared to the GHeP-rog scenario (43.4%, Table 18). There were a disproportionate percentage of GHeP-atyp4a simulations with$d_5^{\bar{3}} = 0$ compared to the GHeP-rog scenario, 19.4% vs. 6.4%, and for the pair of heterogeneous partial agreement parameters $d_5^{\bar{2}}, d_5^{\bar{3}} = 0$ (2.3% vs. 0.5%). Having heterogeneous partial agreement parameters with sufficient statistics equal to zero reduces the number of possible pair-wise comparisons. Consequently, the unadjusted critical p-value is larger and a greater proportion of pair-wise comparisons ($H_o : \hat{d}_5^{\bar{i}} = \hat{d}_5^{\bar{j}}, i \neq j$) will be rejected.

4.2.5. Simulated GHeP-atyp4b Agreement Model Assuming Marginal Homogeneity

Generation of Simulated Tables. One thousand 2^6 contingency tables were generated under the assumption of marginal homogeneity using the parameter estimates for the GHeP-atyp4b model shown in Table 11, column 5. The total number of counts per table ranged from 34 to 115 with a mode of 75 (Table 16). An example of the simulated cell counts for a generated 2^6 contingency tables for this GHeP-atyp4b scenario was presented in Table 12 (col.5).

The rater agreement characteristics across the 1,000 simulated contingency tables are summarized in Table 31. The mean marginal percentages for heterogeneous partial agreement are comparable for Raters 1, 2, 3, 5 and 6 and higher for Rater 4, (~2.3% for the former and ~9.8% for the latter).

Table 31. Marginal Percentages for Different Category Specific Agreement Patterns and Rater Exclusion for the GHeP-atyp4b Model Simulated under the Assumption of Marginal Homogeneity

GHeP-atyp4b Model – Marginal Homogeneity									
Rater *i*	Marginal % for Absence	Marginal % for Presence	G %, d_6	G for Absence %, $d_{6,0}$	G for Presence %, $d_{6,1}$	GP %, d_5	GP for Absence %, $d_{5,0}$	GP for Presence %, $d_{5,1}$	Excluded Rater, % $d_5^{\bar{i}}$
	Mean Marginal % (SE) [min,max]		**Mean Proportion (SE) [min,max]**						
1	50.1 (6.2) [31.0,68.7]	49.8 (6.2) [31.3,69.0]	45.9 (0.27) [21.7,71.0]	23.0 (0.19) [6.3,45.7]	22.9 (0.19) [5.8,42.6]	21.4 (0.22) [4.0,45.9]	10.7 (0.14) [0,28.6]	10.7 (0.14) [0,29.8]	2.3 (0.1) [0,11.2]
2	50.1 (6.2) [32.9,67.7]	49.8 (6.2) [32.3,67.1]							2.2 (0.1) [0,13.3]
3	50.0 (6.0) [26.3,69.8]	50.0 (6.0) [30.2,73.7]							2.2 (0.1) [0,13.1]
4	50.2 (6.1) [31.1,71.0]	49.8 (6.0) [29.0,68.9]							9.8 (0.2) [0,30.5]
5	50.1 (6.2) [32.1,71.4]	49.8 (6.3) [28.6,67.9]							2.3 (0.1) [0,10.4]
6	49.9 (6.3) [30.2,66.7]	50.1 (6.3) [33.3,69.8]							2.4 (0.1) [0,10.9]

* For agreement patterns see Table 4. G= Global agreement; GP= Global and partial agreement

For these GHeP-atyp4b simulated data, only one contingency table had $\hat{d}_5^{\bar{4}} = 0$ (Table 18, row 4, col. 5), and 45.8% of the contingency tables had at least one heterogeneous partial agreement parameter with sufficient statistic equal to zero.

Analysis Assuming Marginal Homogeneity. The number of times each possible pair-wise comparison was statistically significant across the 1,000 simulated contingency tables is summarized in Table 32. Approximately 25% of the unadjusted pair-wise comparisons involving Rater 4 were statistically significant. In contrast, less than 2% of the unadjusted pair-wise comparisons not involving Rater 4 were statistically significant. The only adjusted pair-wise comparisons that were statistically significant involved Rater 4.

Table 32. Number (%) of Times Each Possible Pair-wise Comparison was Statistically Significant Across 1000 Tables Simulated under the GHeP-atyp4b Agreement Model with Marginal Homogeneity

Possible Pair-wise Comparisons (N)	**15 (542)**	**10 (277)**	**6 (126)**	**3 (45)**	**1 (10)**
$\hat{d}_{5}^{\overline{1}}$ vs. $\hat{d}_{5}^{\overline{2}}$					
Unadjusted	2 (0.4)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm’s- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm’s-Sidak	0	0	0	0	--
Missing	0	108	90	43	10
$\hat{d}_{5}^{\overline{1}}$ vs. $\hat{d}_{5}^{\overline{3}}$					
Unadjusted	3 (0.6)	1 (0.6)	0	0	--
Bonferroni	0	0	0	0	--
Holm’s- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm’s-Sidak	0	0	0	0	--
Missing	0	118	98	39	10
$\hat{d}_{5}^{\overline{1}}$ vs. $\hat{d}_{5}^{\overline{4}}$					
Unadjusted	129 (23.8)	68 (31.1)	27 (39.7)	6 (33.3)	0
Bonferroni	5 (0.9)	6 (2.7)	4 (5.9)	1 (5.5)	0
Holm’s- Bonferroni	5 (0.9)	6 (2.7)	4 (5.9)	1 (5.5)	0
Sidak	5 (0.9)	6 (2.7)	4 (5.9)	1 (5.5)	0
Holm’s-Sidak	5 (0.9)	6 (2.7)	4 (5.9)	1 (5.5)	0
Missing	0	58	58	27	9
$\hat{d}_{5}^{\overline{1}}$ vs. $\hat{d}_{5}^{\overline{5}}$					
Unadjusted	0	0	0	0	--
Bonferroni	0	0	0	0	--
Holm’s- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm’s-Sidak	0	0	0	0	--
Missing	0	114	90	36	10

Table 32 (continued)

Possible Pair-wise Comparisons (N)	**15 (542)**	**10 (277)**	**6 (126)**	**3 (45)**	**1 (10)**
$\hat{d}_5^{\overline{1}}$ vs. $\hat{d}_5^{\overline{6}}$					
Unadjusted	2 (0.4)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	107	87	42	10
$\hat{d}_5^{\overline{2}}$ vs. $\hat{d}_5^{\overline{3}}$					
Unadjusted	4 (0.7)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	112	41	39	10
$\hat{d}_5^{\overline{2}}$ vs. $\hat{d}_5^{\overline{4}}$					
Unadjusted	136 (25.1)	74 (32.9)	30 (37.5)	9 (52.9)	0
Bonferroni	5 (0.9)	6 (2.7)	4 (5.0)	2 (11.7)	0
Holm's- Bonferroni	5 (0.9)	6 (2.7)	4 (5.0)	2 (11.7)	0
Sidak	5 (0.9)	6 (2.7)	4 (5.0)	2 (11.7)	0
Holm's-Sidak	5 (0.9)	6 (2.7)	4 (5.0)	2 (11.7)	0
Missing	0	52	46	28	7
$\hat{d}_5^{\overline{2}}$ vs. $\hat{d}_5^{\overline{5}}$					
Unadjusted	2 (0.4)	1 (0.6)	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	108	83	41	10

Table 32 (continued)

Possible Pair-wise Comparisons (N)	**15 (542)**	**10 (277)**	**6 (126)**	**3 (45)**	**1 (10)**
$\hat{d}_{5}^{\overline{2}}$ vs. $\hat{d}_{5}^{\overline{6}}$					
Unadjusted	7 (1.3)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm’s- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm’s-Sidak	0	0	0	0	--
Missing	0	101	43	40	10
$\hat{d}_{5}^{\overline{3}}$ vs. $\hat{d}_{5}^{\overline{4}}$					
Unadjusted	143 (26.3)	67 (31.1)	27 (37.5)	3 (18.8)	0
Bonferroni	2 (0.4)	5 (2.3)	2 (2.7)	1 (6.3)	0
Holm’s- Bonferroni	4 (0.7)	5 (2.3)	2 (2.7)	1 (6.3)	0
Sidak	2 (0.4)	5 (2.3)	2 (2.7)	1 (6.3)	0
Holm’s-Sidak	4 (0.7)	5 (2.3)	2 (2.7)	1 (6.3)	0
Missing	0	62	54	29	9
$\hat{d}_{5}^{\overline{3}}$ vs. $\hat{d}_{5}^{\overline{5}}$					
Unadjusted	4 (0.7)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm’s- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm’s-Sidak	0	0	0	0	--
Missing	0	118	86	43	10
$\hat{d}_{5}^{\overline{3}}$ vs. $\hat{d}_{5}^{\overline{6}}$					
Unadjusted	4 (0.7)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm’s- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm’s-Sidak	0	0	0	0	--
Missing	0	111	91	41	10

Table 32 (continued)

Possible Pair-wise Comparisons (N)	**15 (542)**	**10 (277)**	**6 (126)**	**3 (45)**	**1 (10)**
$\hat{d}_5^{\overline{4}}$ vs. $\hat{d}_5^{\overline{5}}$					
Unadjusted	135 (24.9)	74 (33.8)	31 (39.7)	5 (26.3)	2 (66.7)
Bonferroni	6 (1.1)	6 (2.7)	2 (2.5)	3 (15.8)	2 (66.7)
Holm's- Bonferroni	7 (1.3)	6 (2.7)	2 (2.5)	3 (15.8)	2 (66.7)
Sidak	6 (1.1)	6 (2.7)	2 (2.5)	3 (15.8)	2 (66.7)
Holm's-Sidak	7 (1.3)	6 (2.7)	2 (2.5)	3 (15.8)	2 (66.7)
Missing	0	58	48	26	7
$\hat{d}_5^{\overline{4}}$ vs. $\hat{d}_5^{\overline{6}}$					
Unadjusted	146 (26.9)	66 (29.2)	26 (33.8)	6 (33.3)	0
Bonferroni	5 (0.9)	5 (2.2)	2 (2.6)	0	0
Holm's- Bonferroni	5 (0.9)	5 (2.2)	2 (2.6)	0	0
Sidak	5 (0.9)	5 (2.2)	2 (2.6)	0	0
Holm's-Sidak	5 (0.9)	5 (2.2)	2 (2.6)	0	0
Missing	0	51	49	27	9
$\hat{d}_5^{\overline{5}}$ vs. $\hat{d}_5^{\overline{6}}$					
Unadjusted	6 (1.1)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	107	86	39	10

Table 33 summarizes the power to identify Rater 4 as the atypical rater when $d_5^{\bar{4}}$ differs from one or more of the other $d_5^{\bar{i}}$ ‘s. Using a criterion that $d_5^{\bar{4}}$ differs from at least one $d_5^{\bar{i}}$ (i = 1, 2, 3, 5, or 6), the power to identify Rater 4 as being atypical is 44.2% based on unadjusted comparisons and only 2.9% for the adjusted comparisons.

Table 33. Power to Identify Rater 4 as the Atypical Rater Using Various Criteria By Multiple Comparison Procedure for the GHeP-atyp4b Scenario

Rater 4 Differs from:	Unadjusted	Bonferroni	Holm’s Bonferroni	Sidak	Holm’s Sidak
One rater	0.099	0.007	0.007	0.007	0.5
Two raters	0.114	0.007	0.007	0.007	0.007
Three raters	0.088	0.008	0.006	0.008	0.008
Four raters	0.084	0.006	0.007	0.006	0.005
Five raters	0.057	0.001	0.002	0.001	0.004
At least one rater	0.442	0.029	0.029	0.029	0.029

4.2.6. Simulated G Agreement Model Assuming Marginal Heterogeneity

Generation of Simulated Tables. One thousand 2^6 contingency tables were generated under the assumption of marginal heterogeneity using the parameter estimates for the G model shown in Table 11 (lower half of table, column 2). The total number of counts per table ranged from 37 to 119, with a mode of 74 (Table 34).

Table 34. Descriptive Statistics of Sample Size (Total Counts) of the One Thousand 2^6 Contingency Tables Simulated under the Assumption of Marginal Heterogeneity

	Marginal Heterogeneity				
Scenario	G	GP	GHeP-rog	GHeP-atyp4a	GHeP-atyp4b
Minimum	37	36	32	37	57
Maximum	119	124	143	111	224
Mode	74	65	78	72,74*	110

* Two values indicates a bi-modal distribution

One example of the simulated cell counts of the 64 possible rating patterns for the generated 2^6 contingency tables for the G scenario is presented in Table 35 (col. 2). The shaded patterns represent global agreement or partial agreement. Notice that 20 (83.3%) of the 24 rating patterns representing G agreement were from rating pattern (000000).

Table 35. One Set of Count Data for Five Models Simulated Assuming Marginal Heterogeneity

	Simulated Model				
Rating Pattern	**G**	**GP**	**GHeP-rog**	**GHeP-atyp4a**	**GHeP-atyp4b**
000000	20	32	32	12	23
000001	1	2	2	0	1
000010	0	0	4	2	2
000011	0	0	0	0	0
000100	6	5	4	7	32
000101	4	3	1	1	4
000110	0	1	2	4	8
000111	0	1	0	2	0
001000	0	0	0	2	4
001001	0	0	0	0	0
001010	0	1	0	0	0
001011	1	0	0	0	0
001100	2	3	0	0	8
001101	1	1	0	1	2
001110	0	1	0	0	1
001111	0	1	0	0	0
010000	1	3	2	3	0
010001	1	0	1	0	0
010010	0	1	0	0	0
010011	1	1	0	0	0
010100	4	3	1	4	3
010101	0	0	1	1	0
010110	1	2	0	1	2
010111	0	0	1	0	0
011000	0	0	0	0	0
011001	1	0	0	0	0
011010	1	0	0	0	0
011011	0	0	1	0	0
011100	2	0	1	0	1
011101	1	0	0	2	0
011110	0	0	0	0	0
011111	1	0	2	0	8

Table 35 (continued)

	Simulated Model				
Rating Pattern	**G**	**GP**	**GHeP-rog**	**GHeP-atyp4a**	**GHeP-atyp4b**
100000	0	4	1	2	0
100001	0	0	0	0	0
100010	0	0	0	0	0
100011	0	0	0	0	0
100100	0	2	3	2	0
100101	0	0	1	0	0
100110	0	1	2	0	3
100111	0	1	0	0	0
101000	0	0	0	0	0
101001	0	0	0	0	0
101010	0	1	0	0	0
101011	0	0	0	1	0
101100	2	0	1	1	1
101101	1	1	0	0	0
101110	0	1	0	1	0
101111	0	0	1	1	2
110000	0	0	0	0	0
110001	0	0	1	0	0
110010	0	0	0	1	0
110011	0	0	1	0	0
110100	1	0	2	1	0
110101	1	0	0	0	1
110110	0	1	0	1	0
110111	1	2	3	1	2
111000	0	0	0	0	0
111001	0	0	0	0	0
111010	1	1	0	0	0
111011	0	0	0	0	0
111100	0	1	0	0	1
111101	1	1	5	4	0
111110	0	2	4	4	3
111111	4	7	5	4	9
Sample Size	**61**	**81**	**85**	**66**	**121**

The rater agreement characteristics across the 1,000 simulated contingency tables for the G model are summarized in Table 36. The six raters' mean marginal proportions of rating 'absence' and 'presence' are similar to that observed in the intestinal biopsy example, with Raters 1, 2, 3, 5, and 6 rating 'absence' of the lesion in approximately 72% of the biopsies and

Rater 4 rating ‘absence’ of the lesion in 45.2% of the slides (col. 2). The observed percentage of global agreement is not necessarily divided equally between global agreement on the absence or presence of the lesion (35.8% and 7.1%, respectively). The mean percentage of partial agreement 0.001 was 10.9% (col. 7), of which 50% represented five-way agreement when Rater 4 disagreed with the other raters (col. 10).

Table 36. Marginal Percentages for Different Category Specific Agreement Patterns and Rater Exclusion for the G Agreement Model Simulated under the Assumption of Marginal Heterogeneity

Global Agreement Model – Marginal Heterogeneity									
Rater i	Marginal % for Absence	Marginal % for Presence	G %, d_6	G for Absence %, $d_{6,0}$	G for Presence %, $d_{6,1}$	GP %, d_5	GP for Absence %, $d_{5,0}$	GP for Presence %, $d_{5,1}$	Excluded Rater, % $d_5^{\bar{i}}$
	Mean Marginal % (SD) [min,max]		**Mean Proportion (SE) [min,max]**						
1	77.3 (7.4) [54.1,95.3]	22.7 (7.4) [4.6, 45.9]	42.8 (0.27) [17.6, 78.9]	35.8 (0.25) [14.8, 72.3]	7.1 (0.13) [0, 31.1]	10.9 (0.12) [0, 26.5]	8.8 (0.11) [0,23.1]	2.0 (0.1) [0,11.7]	0.9 (0.03) [0,7.7]
2	68.6 (7.7) [40.2,93.5]	31.3 (7.7) [6.5,59.8]							1.3 (0.04) [0,7.1]
3	71.1 (7.5) [42.9,91.2]	28.8 (7.5) [8.8,57.1]							1.1 (0.04) [0,7.8]
4	45.2 (8.3) [21.5,77.6]	54.8 (8.3) [22.3,78.4]							5.5 (0.1) [0,19.4]
5	72.6 (7.6) [45.5,91.4]	27.4 (7.6) [5.9,54.5]							1.1 (0.04) [0,6.6]
6	74.1 (7.4) [47.9,91.9]	25.9 (7.4) [8.1,52.1]							1.1 (0.04) [0,6.6]

For agreement patterns see Table 4. G= Global agreement; GP= Global and partial agreement

The percentage of the 1,000 simulated 2^6 contingency tables having none, some or all of its heterogeneous partial agreement parameters with sufficient statistics equal to zero is summarized in Table 37. Model G had relatively few (5.1%) simulated tables with no heterogeneous sufficient statistic for a partial agreement parameter equal to zero. This is expected, because under the G model, non-global agreement was spread uniformly across the table rather than being concentrated near the diagonal.

Table 37. Percent of the 1,000 Heterogeneous Simulated 2^6 Contingency Tables with None, Some or All of Its Heterogeneous Partial Agreement Parameters Having Sufficient Statistic Equal to Zero

Sufficient Statistic of GHeP Parameter = 0 in Model $d_5^{\bar{i}} = 0,\ i =$	**G**	**GP**	**GHeP-rog**	**GHeP-atyp4a**	**GHeP-atyp4b**
None	5.1	41.9	36.0	61.1	60.9
6	3.5	8.6	2.8	5.7	5.2
5	3.9	8.4	8.3	6.5	5.9
4	0.1	0	0.3	0.7	0
3	3.5	7.3	20.7	4.6	6.7
2	2.6	4.9	3.1	5.0	5.9
1	5.5	8.2	8.2	4.3	4.6
1, 2	2.7	1.0	0.6	1.2	0.8
1, 3	2.7	1.6	5.0	0.6	0.8
1, 4	0.2	0	0	0	0
1, 5	3.7	1.6	1.7	0.8	0.8
1, 6	2.7	1.3	0.6	1.2	0.9
2, 3	2.5	1.3	1.9	0.6	1.0
2, 4	0.2	0	0	0.1	0
2, 5	2.0	1.1	0.7	1.2	0.6
2, 6	2.7	1.6	0.2	1.7	0.7
3, 4	0.4	0	0.4	0.1	0
3, 5	2.6	1.5	4.2	0.7	0.9
3, 6	3.3	1.7	1.5	1.1	1.2
4, 5	0.1	0	0	0	0
5, 6	0.3	0	0	0	0

Table 37 (continued)

Sufficient Statistic of GHeP Parameter = 0 in Model $d_5^{\overline{i}} = 0,\ i =$	**G**	**GP**	**GHeP-rog**	**GHeP-atyp4a**	**GHeP-atyp4b**
1, 2, 3	2.6	2.6	0.5	1.2	1.1
1, 2, 4	2.6	0.5	0.6	0	0.3
1, 2, 5	0.1	0	0	0	0
1, 2, 6	2.5	0.6	0.2	0.2	0.1
2, 3, 4	2.5	0.2	0.1	0.1	0.2
2, 3, 5	0.1	0	0	0	0
2, 3, 6	2.7	0.2	0.5	0.3	0
3, 4, 5	2.3	0.2	0.1	0.4	0.3
3, 4, 6	0	0	0	0	0
1, 3, 4	0.1	0	0	0.1	0
1, 3, 5	0.1	0	0	0	0
1, 3, 6	4.0	0.3	0.7	0	0.3
1, 4, 5	2.8	0.5	0.3	0	0
1, 5, 6	0	0	0	0	0
1, 4, 6	2.9	0.3	0.2	0	0.2
2, 4, 5	0.1	0	0	0	0
2, 4, 6	0	0	0	0	0
4, 5, 6	0	0	0	0	0
3, 5, 6	0.1	0	0	0	0
2, 5, 6	2.9	0.7	0.1	0.2	0.3
3, 4, 5, 6	2.6	0.6	0.2	0.1	0
2, 4, 5, 6	0.1	0	0	0	0
2, 3, 5, 6	0.1	0	0	0	0
2, 3, 4, 6	2.3	0.2	0.1	0	0.1
2, 3, 4, 6	0	0	0	0	0
1, 4, 5, 6	0	0	0	0	0
1, 3, 5, 6	0	0	0	0	0
1, 3, 4, 6	4.3	0.2	0	0	0
1, 3, 4, 5	0.1	0	0	0	0
1, 2, 5, 6	0.1	0	0	0	0
1, 2, 4, 6	2.5	0.3	0	0	0
1, 2, 4, 6	0.1	0	0	0	0
1, 2, 3, 5	3.1	0	0	0	0
1, 4, 5, 6	0	0	0	0	0
1, 3, 5, 6	0	0	0	0	0
1, 3, 4, 6	4.3	0.2	0	0	0
1, 3, 4, 5	0.1	0	0	0	0

Table 37 (continued)

Sufficient Statistic of GHeP Parameter = 0 in Model $d_5^{\,i} = 0$	**G**	**GP**	**GHeP-rog**	**GHeP-atyp4a**	**GHeP-atyp4b**
1, 2, 5, 6	0.1	0	0	0	0
1, 2, 4, 6	2.5	0.3	0	0	0
1, 2, 4, 6	0.1	0	0	0	0
1, 2, 3, 5	3.1	0	0	0	0
1, 2, 3, 4	2.2	0	0.1	0.2	0.1
2, 3, 4, 5, 6	0.1	0.4	0.1	0	0.1
1, 3, 4, 5, 6	0.1	0	0	0	0
1, 2, 4, 5, 6	0.2	0	0	0	0
1, 3, 4, 5, 6	0.1	0	0	0	0
1, 2, 3, 5, 6	2.9	0	0	0	0
1, 2, 3, 4, 6	0.1	0	0	0	0
1, 2, 3, 4, 5	0.2	0.2	0	0	0
All	0.1	0	0	0	0
TOTAL	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**

The sufficient statistic for the heterogeneous partial agreement parameter $d_5^{\bar{4}}$ was zero in one (0.1%) of the 1,000 simulated contingency tables (shaded row, col. 2). Because each set of 1,000 simulated 2^6 contingency tables included tables where some GHeP parameters had sufficient statistics equal to zero, the actual number of possible pair-wise comparisons was less than 1,000.

Analysis Assuming Marginal Homogeneity. The number of times each possible pair-wise comparison was statistically significant across the 1,000 simulated contingency tables is summarized in Table 38. All of the statistically significant unadjusted pair-wise comparisons involved Rater 4. None of the adjusted comparisons $\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{4}}$ was statistically significant. The percentage of significant adjusted pair-wise comparisons ranged from 0.2% ($d_5^{\bar{4}}$ vs. $d_5^{\bar{6}}$) to 0.6% ($d_5^{\bar{4}}$ vs. $d_5^{\bar{5}}$).

Table 38. Number (%) of Times Each Possible Pair-wise Comparison was Statistically Significant Across 1000 Tables Simulated under the G Model with Marginal Heterogeneity when the Data were Analyzed Assuming Marginal Homogeneity

Comparison	Unadjusted	Bonferroni	Holm's - Bonferroni	Sidak	Holm's - Sidak
	n (%)				
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{2}}$	0	0	0	0	0
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{3}}$	0	0	0	0	0
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{4}}$	15(3.2)	0	0	0	0
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{5}}$	0	0	0	0	0
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{6}}$	0	0	0	0	0
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{3}}$	0	0	0	0	0
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{4}}$	20(3.6)	2(0.3)	2(0.3)	2(0.3)	2(0.3)
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{5}}$	0	0	0	0	0
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{6}}$	0	0	0	0	0
$\hat{d}_5^{\bar{3}}$ vs. $\hat{d}_5^{\bar{4}}$	27(5.4)	2(0.4)	2(0.4)	2(0.4)	2(0.4)
$\hat{d}_5^{\bar{3}}$ vs. $\hat{d}_5^{\bar{5}}$	0	0	0	0	0
$\hat{d}_5^{\bar{3}}$ vs. $\hat{d}_5^{\bar{6}}$	0	0	0	0	0
$\hat{d}_5^{\bar{4}}$ vs. $\hat{d}_5^{\bar{5}}$	28(5.6)	3(0.6)	3(0.6)	3(0.6)	3(0.6)
$\hat{d}_5^{\bar{4}}$ vs. $\hat{d}_5^{\bar{6}}$	24(4.7)	1(0.2)	1(0.2)	1(0.2)	1(0.2)
$\hat{d}_5^{\bar{5}}$ vs. $\hat{d}_5^{\bar{6}}$	0	0	0	0	0

The Type I Error to identify Rater 4 as the atypical rater when $d_5^{\bar{4}}$ differs from only one $d_5^{\bar{i}}$ (i = 1, 2, 3, 5, or 6) or more than one $d_5^{\bar{i}}$ is summarized in Table 39.

The Type I Error is 6.6% when no adjustment is made for the number of comparisons and 0.7% when adjusted using each multiple comparison procedure considered.

Table 39. Type I Error to Identify Rater 4 as the Atypical Rater for G Scenario Simulated under Marginal Heterogeneity when the Data were Analyzed Assuming Marginal Homogeneity

Rater 4 Differs from:	Unadjusted	Bonferroni	Holm's-Bonferroni	Sidak	Holm's-Sidak
One rater	0.026	0.006	0.006	0.006	0.006
Two raters	0.025	0.001	0.001	0.001	0.001
Three raters	0.013	0	0	0	0
Four raters	0.002	0	0	0	0
Five raters	0	0	0	0	0
At least one rater	0.066	0.07	0.07	0.07	0.07

Analysis Assuming Marginal Heterogeneity. The number of times each possible pair-wise comparison was statistically significant across the 1,000 simulated contingency tables is summarized in Table 40. The only significant pair-wise comparisons, unadjusted or adjusted, involved $d_5^{\bar{4}}$.

Table 40. Number (%) of Times Each Possible Pair-wise Comparison was Statistically Significant Across 1000 Tables Simulated under the G Model with Marginal Heterogeneity when the Data were Analyzed Assuming Marginal Heterogeneity

Comparison	Unadjusted	Bonferroni	Holm's - Bonferroni	Sidak	Holm's - Sidak
	n (%)				
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{2}}$	0	0	0	0	0
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{3}}$	0	0	0	0	0
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{4}}$	18(3.8)	3(0.6)	3(0.6)	3(0.6)	3(0.6)
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{5}}$	0	0	0	0	0
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{6}}$	0	0	0	0	0
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{3}}$	0	0	0	0	0
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{4}}$	37(6.6)	6(1.1)	6(1.1)	7(1.3)	7(1.3)

Table 40 (continued)

Comparison	Unadjusted	Bonferroni	Holm's - Bonferroni	Sidak	Holm's - Sidak
	n (%)				
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{5}}$	0	0	0	0	0
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{6}}$	0	0	0	0	0
$\hat{d}_5^{\bar{3}}$ vs. $\hat{d}_5^{\bar{4}}$	27(5.4)	10(2.0)	10(2.0)	10(2.0)	10(2.0)
$\hat{d}_5^{\bar{3}}$ vs. $\hat{d}_5^{\bar{5}}$	0	0	0	0	0
$\hat{d}_5^{\bar{3}}$ vs. $\hat{d}_5^{\bar{6}}$	0	0	0	0	0
$\hat{d}_5^{\bar{4}}$ vs. $\hat{d}_5^{\bar{5}}$	27(5.4)	5(1.0)	5(1.0)	5(1.0)	5(1.0)
$\hat{d}_5^{\bar{4}}$ vs. $\hat{d}_5^{\bar{6}}$	27(5.3)	5(0.9)	6(1.1)	5(0.9)	6(1.1)
$\hat{d}_5^{\bar{5}}$ vs. $\hat{d}_5^{\bar{6}}$	0	0	0	0	0

The Type I Error to identify Rater 4 as the atypical rater when $d_5^{\bar{4}}$ differs from one or more of the other $d_5^{\bar{i}}$ is summarized in Table 41. The overall Type I Error is 11% when no adjustment is made for the number of comparisons and ~ 3.0% when each of the multiple comparison procedures is used.

Table 41. Type I Error to Identify Rater 4 as the Atypical Rater for G Scenario Simulated under Marginal Heterogeneity when the Data were Analyzed Assuming Marginal Heterogeneity

	Multiple Comparison Procedure				
Rater 4 Differs from:	**Unadjusted**	**Bonferroni**	**Holm's - Bonferroni**	**Sidak**	**Holm's - Sidak**
One rater	0.083	0.026	0.026	0.028	0.027
Two raters	0.022	0.002	0.003	0.002	0.003
Three raters	0.004	0	0	0	0
Four raters	0.001	0	0	0	0
Five raters	0	0	0	0	0
At least one rater	0.11	0.028	0.029	0.03	0.03

* 0.06% not evaluable

4.2.7. Simulated GP Agreement Model Assuming Marginal Heterogeneity

Generation of Simulated Tables. One thousand 2^6 contingency tables were generated under the assumption of marginal heterogeneity using the parameter estimates for the GP model shown in Table 11, column 3. The total number of counts per table ranged from 36 to 124 (Table 34). One example of the simulated cell counts of the 64 possible rating patterns for the generated 2^6 contingency tables was presented in Table 35 (col. 2). Thirty-two (~ 82%) of the 39 rating patterns representing global agreement were from rating pattern (000000) and ~ 32% of the rating patterns represented partial agreement.

The rater agreement characteristics across the 1,000 simulated contingency tables for the GA model are summarized in Table 42. The six raters' mean marginal proportions of rating 'absence' and 'presence' from the GP model simulated under the assumption of marginal heterogeneity are similar to those observed in the intestinal biopsy data, as are the mean percentages of global and partial agreement (global agreement; 42.3% vs. 44.1%, partial agreement; 25.5% vs. 25.0%). Relatively more GP agreement occurred for the absence (20.4%) than the presence (5.1%) of the lesion. The mean percentages of five-way agreement when Raters 1, 2, 3, 5, and 6 were excluded were similar (~2.6%) and less than the five-way agreement when Rater 4 was excluded (12.5%, col 10.)

Approximately 42% of the simulations using the GP model had no sufficient statistic equal to zero for a heterogeneous partial agreement parameter. No sufficient statistics for $d_{5}^{\bar{4}}$ were zero (row 4, col. 3, Table 37).

Table 42. Marginal Percentages for Different Category Specific Agreement Patterns and Rater Exclusion for the GP Agreement Model Simulated under the Assumption of Marginal Heterogeneity

Global & Partial Agreement Model – Heterogeneity										
Rater *i*	Marginal % for Absence	Marginal % for Presence	G %, d_6	G for Absence %, $d_{6,0}$	G for Presence %, $d_{6,1}$	GP %, d_5	GP for Absence %, $d_{5,0}$	GP for Presence %, $d_{5,1}$	Excluded Rater, % $d_5^{\bar{i}}$
	Mean Marginal % (SE) [min,max]		**Mean Proportion (SE) [min,max]**							
1	77.3 (6.9) [47.7,96.8]	22.6 (6.9) [3.2,52.2]	42.3 (0.27) [17.3,74.3]	33.7 (0.24) [9.6,62.0]	8.6 (0.14) [0,28.4]	25.5 (0.23) [6.5,47.2]	20.4 (0.20) [4.2,41.5]	5.1 (0.1) [0,15.9]	2.7 (0.1) [0,15.0]
2	68.2 (7.8) [39.7,91.3]	31.8 (7.8) [8.83,60.3]							2.9 (0.1) [0,16.7]
3	71.2 (7.5) [49.3,92.3]	28.8 (7.5) [7.7,50.7]							2.6 (0.1) [0,9.8]
4	44.9 (8.2) [20.0,73.8]	55.0 (8.2) [26.1,80.0]							12.5 (0.2) [1.1,31.5]
5	72.6 (7.5) [40.0,91.3]	27.4 (7.5) [8.8,60.0]							2.4 (0.1) [0,11.3]
6	74.0 (7.1) [52.5,93.1]	72.6 (7.1) [6.9,47.5]							2.4 (0.1) [0,10.1]

For agreement patterns see Table 4. G= Global agreement; GP= Global and partial agreement

Analysis Assuming Marginal Homogeneity. The number of times each possible pair-wise comparison was statistically significant across the 1,000 simulated contingency tables is summarized in Table 43. The vast majority of the significant unadjusted pair-wise comparisons and all of the significant adjusted comparisons involved $\hat{d}_5^{\bar{4}}$.

Table 43. Number (%) of Times Each Possible Pair-wise Comparison was Statistically Significant Across 1000 Tables Simulated under the GP Model with Marginal Heterogeneity when Data were Analyzed Assuming Marginal Homogeneity

Comparison	Unadjusted	Bonferroni	Holm's - Bonferroni	Sidak	Holm's - Sidak
	n (%)				
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{2}}$	5 (0.7)	0	0	0	0
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{3}}$	2 (0.3)	0	0	0	0
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{4}}$	358 (43.2)	29 (3.5)	30 (3.6)	29 (3.5)	30 (3.6)
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{5}}$	2 (0.3)	0	0	0	0
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{6}}$	1 (0.2)	0	0	0	0
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{3}}$	2 (0.3)	0	0	0	0
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{4}}$	365 (42.1)	24 (2.8)	27 (3.1)	24 (2.8)	29 (3.3)
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{5}}$	2 (0.2)	0	0	0	0
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{6}}$	4 (0.5)	0	0	0	0
$\hat{d}_5^{\bar{3}}$ vs. $\hat{d}_5^{\bar{4}}$	357 (42.9)	26 (3.1)	28 (3.4)	26 (3.1)	28 (3.4)
$\hat{d}_5^{\bar{3}}$ vs. $\hat{d}_5^{\bar{5}}$	0	0	0	0	0
$\hat{d}_5^{\bar{3}}$ vs. $\hat{d}_5^{\bar{6}}$	1 (0.2)	0	0	0	0
$\hat{d}_5^{\bar{4}}$ vs. $\hat{d}_5^{\bar{5}}$	334 (41.3)	25 (3.1)	29 (3.6)	25 (3.1)	29 (3.6)
$\hat{d}_5^{\bar{4}}$ vs. $\hat{d}_5^{\bar{6}}$	346 (42.8)	25 (3.1)	27 (3.3)	25 (3.1)	27 (3.3)
$\hat{d}_5^{\bar{5}}$ vs. $\hat{d}_5^{\bar{6}}$	0	0	0	0	0

The Type I Error to identify Rater 4 as the atypical rater when $d_5^{\bar{4}}$ differs from one or more one of the other $d_5^{\bar{i}}$ is summarized in Table 44. The Type I Error is 58.8% without adjusting for the number of comparisons and 4.3% when adjustments are made.

Table 44. Type I Error to Identify Rater 4 as the Atypical Rater for GP Scenario Simulated under Marginal Heterogeneity when the Data were Analyzed Assuming Marginal Homogeneity

Rater 4 Differs from:	Unadjusted	Bonferroni	Holm's - Bonferroni	Sidak	Holm's - Sidak
One rater	0.088	0.007	0.005	0.007	0.005
Two raters	0.09	0.009	0.008	0.009	0.008
Three raters	0.129	0.008	0.006	0.008	0.006
Four raters	0.15	0.015	0.018	0.015	0.016
Five raters	0.101	0.004	0.006	0.004	0.008
At least one rater	0.588	0.043	0.043	0.043	0.043

* 0.06% not evaluable

Analysis Assuming Marginal Heterogeneity. The number of times each possible pair-wise comparison was statistically significant across the 1,000 simulated contingency tables is summarized in Table 45. In contrast to the results from the analysis assuming marginal homogeneity, the percentage of significant unadjusted pair-wise comparisons ranged from 0% ($\hat{d}_5^{\bar{3}}$ vs. $\hat{d}_5^{\bar{5}}$) to 3.8% ($\hat{d}_5^{\bar{4}}$ vs. $\hat{d}_5^{\bar{5}}$). After adjusting for the number of multiple comparisons, only the comparisons involving $\hat{d}_5^{\bar{4}}$ remained significant, and the percentage of significant pair-wise comparisons ranged from 0.1% to 0.7%.

Table 45. Number (%) of Times Each Possible Pair-wise Comparison was Statistically Significant Across 1000 Tables Simulated under the GP Agreement Model Assuming Marginal Heterogeneity when the Data were Analyzed Assuming Marginal Heterogeneity

Comparison	Unadjusted	Bonferroni	Holm's - Bonferroni	Sidak	Holm's - Sidak
	n (%)				
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{2}}$	3 (0.4)	0	0	0	0
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{3}}$	3 (0.4)	0	0	0	0
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{4}}$	28 (3.3)	4 (0.4)	4 (0.4)	4 (0.4)	4 (0.4)
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{5}}$	3 (0.4)	0	0	0	0
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{6}}$	5 (0.8)	0	0	0	0
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{3}}$	5 (0.7)	0	0	0	0
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{4}}$	26 (3.0)	1 (0.1)	2 (0.2)	1 (0.1)	2 (0.2)
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{5}}$	4 (0.5)	0	0	0	0
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{6}}$	5 (0.7)	0	0	0	0
$\hat{d}_5^{\bar{3}}$ vs. $\hat{d}_5^{\bar{4}}$	27 (3.3)	5 (0.6)	5 (0.6)	5 (0.6)	5 (0.6)
$\hat{d}_5^{\bar{3}}$ vs. $\hat{d}_5^{\bar{5}}$	0	0	0	0	0
$\hat{d}_5^{\bar{3}}$ vs. $\hat{d}_5^{\bar{6}}$	2 (0.3)	0	0	0	0
$\hat{d}_5^{\bar{4}}$ vs. $\hat{d}_5^{\bar{5}}$	31 (3.8)	2 (0.3)	2 (0.3)	2 (0.3)	2 (0.3)
$\hat{d}_5^{\bar{4}}$ vs. $\hat{d}_5^{\bar{6}}$	25 (3.1)	5 (0.6)	6 (0.7)	5 (0.6)	6 (0.7)
$\hat{d}_5^{\bar{5}}$ vs. $\hat{d}_5^{\bar{6}}$	4 (0.6)	0	0	0	0

The Type I Error to identify Rater 4 as the atypical rater when $d_5^{\bar{4}}$ differs from one or more of the other $d_5^{\bar{i}}$ is summarized in Table 46. The unadjusted Type I Error is 9.6%, compared to 1.5% when adjusted for the number of pair-wise comparisons.

Table 46. Type I Error to Identify Rater 4 as the Atypical Rater for GP Scenario Simulated under Marginal Heterogene ity when the Data were Analyzed Assuming Marginal Heterogeneity

	Multiple Comparison Procedure				
Rater 4 Differs from:	**Unadjusted**	**Bonferroni**	**Holm's - Bonferroni**	**Sidak**	**Holm's - Sidak**
One rater	0.071	0.013	0.013	0.013	0.013
Two raters	0.014	0.002	0.001	0.002	0.001
Three raters	0.007	0	0	0	0
Four raters	0.003	0	0.001	0	0.001
Five raters	0.001	0	0	0	0
At least one rater	0.096	0.015	0.015	0.015	0.015

4.2.8. Simulated GHeP-rog Agreement Model Assuming Marginal Heterogeneity

Generation of Simulated Tables. One thousand 2^6 contingency tables were generated under the assumptio n of marginal heterogeneity using the parameter estimates for the GHeP-rog model shown in Table 11, column 4. The total number of counts ranged from 32 to 143 (Table 34). One example of the simulated cell counts of the 64 possible rating patterns for the generated 2^6 contingency tables for the GHeP-rog agreement model was presented in Table 35 (col. 3). Thirty-two of the 37 (~86%) rating patterns representing global agreement were from rating pattern (000000) and ~33% of the rating patterns represented partial agreement.

The rater agreement characteristics across the 1,000 simulated contingency tables for the GHeP-rog agreement model are summarized in Table 47. The six raters' mean marginal proportions of rating 'absence' and 'presence' are similar to that observed in the intestinal biopsy

Table 47. Marginal Percentages for Different Category Specific Agreement Patterns and Rater Exclusion for the GHeP-rog Agreement Model Simulated under the Assumption of Marginal Heterogeneity

Global & Partial Agreement Model – Heterogeneity									
Rater *i*	Marginal % for Absence	Marginal % for Presence	G %, d_6	G for Absence %, $d_{6,0}$	G for Presence %, $d_{6,1}$	GP %, d_5	GP for Absence %, $d_{5,0}$	GP for Presence %, $d_{5,1}$	Excluded Rater, % $d_5^{\bar{i}}$
	Mean Marginal % (SE) [min,max]		**Mean Proportion (SE) [min,max]**						
1	76.6 (7.4) [46.4,96.8]	23.4 (7.4) [3.2,53.6]	40.6 (0.25) [18.2,63.3]	33.3 (0.24) [10.3,60.2]	7.3 (0.12) [0,23.2]	28.2 (0.25) [9.9,56.5]	19.6 (0.20) [1.7,46.3]	8.6 (0.15) [0,36.4]	3.7 (0.1) [0,29.2]
2	67.4 (8.0) [42.6,96.7]	32.6 (8.0) [3.3,57.4]							4.7 (0.1) [0,27.5]
3	70.3 (7.8) [44.4,89.3]	29.7 (7.8) [10.7,55.6]							2.2 (0.1) [0,23.8]
4	44.4 (8.5) [18.9,71.9]	55.6 (8.5) [28.1,81.1]							8.7 (0.2) [0,39.1]
5	71.6 (7.6) [45.7,92.8]	28.4 (7.6) [7.2,54.3]							3.6 (0.1) [0,31.4]
6	73.1 (7.6) [48.8,93.2]	26.9 (7.6) [6.8,51.2]							5.2 (0.1) [0,29.5]

* For agreement patterns see Table 4. G= Global agreement; GP= Global and partial agreement

data. The mean percentages of global and partial agreement were also comparable to that seen in the intestinal biopsy data (global agreement; 40.6% vs. 44.1%, partial agreement; 28.2% vs. 25.0%). In contrast to that observed in the intestinal data, the mean percentage partial agreement for absence of the lesion was greater than that for presence of the lesion in the simulated data (19.6 % vs. 5.1%).

Approximately 35% of the simulations using the GHeP-rog model had no sufficient statistic for a heterogeneous partial agreement parameter equal to zero. Only four contingency tables had a sufficient statistic for$d_5^{\bar{4}}$ equal to zero (row 4, col. 4).

Analysis Assuming Marginal Homogeneity. The number of times each possible pair-wise comparison was statistically significant across the 1,000 simulated contingency tables is summarized in Table 48. Relatively more significant unadjusted pair-wise comparisons involved Rater 4.

Table 48. Number (%) of Times Each Possible Pair-wise Comparison was Statistically Significant Across 1000 Tables Simulated under the GHeP-Rog Agreement Model Assuming Marginal Heterogeneity when Data were Analyzed Assuming Marginal Homogeneity

Possible Pair-wise Comparisons (N)	**15 (360)**	**10 (434)**	**6 (173)**	**3 (173)**	**1 (2)**
$\hat{d}_5^{\overline{1}}$ vs. $\hat{d}_5^{\overline{2}}$					
Unadjusted	37 (10.3)	25 (7.8)	4 (6.0)	4 (6.0)	--
Bonferroni	1 (0.2)	2 (0.6)	0	0	--
Holm's- Bonferroni	1 (0.2)	2 (0.6)	0	0	--
Sidak	1 (0.2)	2 (0.6)	0	0	--
Holm's-Sidak	1 (0.2)	2 (0.6)	0	0	--
Missing	0	113	107	107	2
$\hat{d}_5^{\overline{1}}$ vs. $\hat{d}_5^{\overline{3}}$					
Unadjusted	29 (8.0)	13 (9.0)	2 (14.3)	2 (14.3)	--
Bonferroni	1 (0.2)	0	0	0	--
Holm's- Bonferroni	1 (0.2)	0	0	0	--
Sidak	1 (0.2)	0	0	0	--
Holm's-Sidak	1 (0.2)	0	0	0	--
Missing	0	289	159	159	2
$\hat{d}_5^{\overline{1}}$ vs. $\hat{d}_5^{\overline{4}}$					
Unadjusted	56 (15.5)	64 (18.3)	13 (14.4)	13 (14.4)	--
Bonferroni	4 (1.1)	6 (1.7)	1 (1.1)	1 (1.1)	--
Holm's- Bonferroni	4 (1.1)	7 (2.0)	1 (1.1)	1 (1.1)	--
Sidak	4 (1.1)	6 (1.7)	1 (1.1)	1 (1.1)	--
Holm's-Sidak	4 (1.1)	7 (2.0)	1 (1.1)	1 (1.1)	--
Missing	0	85	83	83	2
$\hat{d}_5^{\overline{1}}$ vs. $\hat{d}_5^{\overline{5}}$					
Unadjusted	23 (6.4)	18 (6.7)	2 (5.0)	2 (5.0)	--
Bonferroni	2 (0.5)	2 (0.7)	1 (2.5)	1 (2.5)	--
Holm's- Bonferroni	2 (0.5)	2 (0.7)	1 (2.5)	1 (2.5)	--
Sidak	2 (0.5)	2 (0.7)	1 (2.5)	1 (2.5)	--
Holm's-Sidak	2 (0.5)	2 (0.7)	1 (2.5)	1 (2.5)	--
Missing	0	165	133	133	2

Table 48 (continued)

Possible Pair-wise Comparisons	**15**	**10**	**6**	**3**	**1**
(N)	**(360)**	**(434)**	**(173)**	**(173)**	**(2)**
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{6}}$					
Unadjusted	31 (8.6)	30 (9.2)	5 (6.9)	5 (6.9)	--
Bonferroni	1 (0.2)	3 (0.9)	3 (4.2)	3 (4.2)	--
Holm's- Bonferroni	1 (0.2)	3 (0.9)	3 (4.2)	3 (4.2)	--
Sidak	1 (0.2)	3 (0.9)	3 (4.2)	3 (4.2)	--
Holm's-Sidak	1 (0.2)	3 (0.9)	3 (4.2)	3 (4.2)	--
Missing	0	110	101	101	2
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{3}}$					
Unadjusted	34 (9.4)	20 (10.2)	6 (21.4)	6 (21.4)	--
Bonferroni	0	2 (1.0)	1 (3.6)	1 (3.6)	--
Holm's- Bonferroni	0	2 (1.0)	1 (3.6)	1 (3.6)	--
Sidak	0	2 (1.0)	1 (3.6)	1 (3.6)	--
Holm's-Sidak	1 (0.2)	2 (1.0)	1 (3.6)	1 (3.6)	--
Missing	0	238	145	145	2
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{4}}$					
Unadjusted	46 (12.7)	71 (17.8)	23 (17.0)	23 (17.0)	--
Bonferroni	5 (1.4)	6 (1.5)	4 (2.9)	4 (3.0)	--
Holm's- Bonferroni	5 (1.4)	6 (1.5)	4 (2.9)	4 (3.0)	--
Sidak	5 (1.4)	6 (1.5)	4 (2.9)	4 (3.0)	--
Holm's-Sidak	5 (1.4)	8 (2.0)	4 (2.9)	4 (3.0)	--
Missing	0	34	38	38	2
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{5}}$					
Unadjusted	36 (10.0)	25 (7.8)	5 (6.7)	5 (6.7)	--
Bonferroni	1 (0.2)	3 (0.9)	1 (1.3)	1 (1.3)	--
Holm's- Bonferroni	1 (0.2)	3 (0.9)	1 (1.3)	1 (1.3)	--
Sidak	1 (0.2)	3 (0.9)	1 (1.3)	1 (1.3)	--
Holm's-Sidak	1 (0.2)	3 (0.9)	1 (1.3)	1 (1.3)	--
Missing	0	114	98	98	2

Table 48 (continued)

Possible Pair-wise Comparisons (N)	**15 (360)**	**10 (434)**	**6 (173)**	**3 (173)**	**1 (2)**
$\hat{d}_5^{\overline{2}}$ vs. $\hat{d}_5^{\overline{6}}$					
Unadjusted	30 (8.3)	29 (7.7)	11 (9.7)	11 (9.7)	--
Bonferroni	1 (0.2)	2 (0.5)	5 (4.4)	5 (4.4)	--
Holm's- Bonferroni	1 (0.2)	2 (0.5)	5 (4.4)	5 (4.4)	--
Sidak	1 (0.2)	2 (0.5)	5 (4.4)	5 (4.4)	--
Holm's-Sidak	1 (0.2)	2 (0.5)	5 (4.4)	5 (4.4)	--
Missing	0	59	60	60	2
$\hat{d}_5^{\overline{3}}$ vs. $\hat{d}_5^{\overline{4}}$					
Unadjusted	81 (22.5)	56 (25.0)	10 (23.2)	10 (23.2)	--
Bonferroni	5 (1.4)	4 (1.8)	1 (2.3)	1 (2.3)	--
Holm's- Bonferroni	6 (1.7)	4 (1.8)	1 (2.3)	1 (2.3)	--
Sidak	5 (1.4)	4 (1.8)	1 (2.3)	1 (2.3)	--
Holm's-Sidak	6 (1.7)	6 (2.6)	1 (2.3)	1 (2.3)	--
Missing	0	210	130	130	2
$\hat{d}_5^{\overline{3}}$ vs. $\hat{d}_5^{\overline{5}}$					
Unadjusted	23 (6.4)	9 (6.3)	2 (14.3)	2 (14.3)	--
Bonferroni	3 (0.8)	2 (1.4)	0	0	--
Holm's- Bonferroni	3 (0.8)	2 (1.4)	0	0	--
Sidak	3 (0.8)	2 (1.4)	0	0	--
Holm's-Sidak	3 (0.8)	2 (1.4)	0	0	--
Missing	0	290	159	159	2
$\hat{d}_5^{\overline{3}}$ vs. $\hat{d}_5^{\overline{6}}$					
Unadjusted	29 (8.1)	21 (10.6)	2 (6.7)	2 (6.7)	--
Bonferroni	1 (0.2)	0	2 (6.7)	2 (6.7)	--
Holm's- Bonferroni	1 (0.2)	1 (0.5)	2 (6.7)	2 (6.7)	--
Sidak	1 (0.2)	0	2 (6.7)	2 (6.7)	--
Holm's-Sidak	1 (0.2)	1 (0.5)	2 (6.7)	2 (6.7)	--
Missing	0	235	143	143	2

Table 48 (continued)

Possible Pair-wise Comparisons (N)	**15 (360)**	**10 (434)**	**6 (173)**	**3 (173)**	**1 (2)**
$\hat{d}_5^{\overline{4}}$ vs. $\hat{d}_5^{\overline{5}}$					
Unadjusted	71 (19.7)	62 (17.8)	23 (23.4)	23 (23.5)	0
Bonferroni	4 (1.1)	6 (1.7)	3 (3.1)	3 (3.0)	0
Holm's- Bonferroni	4 (1.1)	7 (2.0)	3 (3.1)	3 (3.0)	0
Sidak	4 (1.1)	6 (1.7)	3 (3.1)	3 (3.0)	0
Holm's-Sidak	4 (1.1)	9 (2.6)	3 (3.1)	3 (3.0)	0
Missing	0	86	75	75	1
$\hat{d}_5^{\overline{4}}$ vs. $\hat{d}_5^{\overline{6}}$					
Unadjusted	59 (16.4)	72 (17.8)	24 (17.0)	24 (17.0)	0
Bonferroni	2 (0.5)	11 (2.7)	4 (2.8)	4 (2.8)	0
Holm's- Bonferroni	3 (0.8)	12 (3.0)	4 (2.8)	4 (2.8)	0
Sidak	2 (0.5)	11 (2.7)	4 (2.8)	4 (2.8)	0
Holm's-Sidak	3 (0.8)	12 (3.0)	4 (2.8)	4 (2.8)	0
Missing	0	31	32	32	1
$\hat{d}_5^{\overline{5}}$ vs. $\hat{d}_5^{\overline{6}}$					
Unadjusted	25 (6.9)	27 (8.3)	10 (12.6)	10 (12.6)	--
Bonferroni	1 (0.2)	4 (1.3)	2 (2.5)	2 (2.5)	--
Holm's- Bonferroni	1 (0.2)	4 (1.3)	2 (2.5)	2 (2.5)	--
Sidak	1 (0.2)	4 (1.3)	2 (2.5)	2 (2.5)	--
Holm's-Sidak	1 (0.2)	4 (1.3)	2 (2.5)	2 (2.5)	--
Missing	0	111	94	94	2

The power to identify Rater 4 as the atypical rater when $\boldsymbol{d}_5^{\bar{4}}$ differs from one or more of the other $\boldsymbol{d}_5^{\bar{i}}$ is summarized in Table 49.

Table 49. Power to Identify Rater 4 as the Atypical Rater Using Various Criteria By Multiple Comparison Procedure for GHeP-rog Scenario, Simulated Assuming Marginal Heterogeneity and when the Data were Analyzed Assuming Marginal Homogeneity

Rater 4 Differs from:	Unadjusted	Bonferroni	Holm's - Bonferroni	Sidak	Holm's-Sidak
One rater	0.148	0.014	0.014	0.014	0.011
Two raters	0.080	0.006	0.006	0.006	0.006
Three raters	0.059	0.004	0.004	0.004	0.007
Four raters	0.048	0.005	0.005	0.005	0.005
Five raters	0.017	0.002	0.003	0.003	0.003
At least one rater	0.352	0.031	0.032	0.031	0.032

The power of the approach was 35.2% for the unadjusted pair-wise comparisons. Using any four of the multiple comparison procedures considered, the power was reduced to ~ 3.2%. A rater other than Rater 4 was identified as the atypical rater in ~26% of the simulations based upon unadjusted p-values and in ~ 3% based on adjusted p-values (Table 50).

Table 50. Proportion (%) of Simulations that Identify a Rater Other Than Rater 4 as the Atypical Rater by Multiple Comparison Procedure for Scenarios Simulated under the Assumption of Marginal Heterogeneity and when the Data were Analyzed Assuming Marginal Homogeneity

At least one rater is different	Unadjusted	Bonferroni	Holm's Bonferroni	Sidak	Holm's Sidak
GHeP-rog	25.7	2.42	2.72	2.42	2.72
GHeP-atyp4a	2.6	0	0	0	0
GHeP-atyp4b	2.3	0	0	0	0

The overall probability that any rater is identified as an atypical rater is approximately 12% based on unadjusted comparisons and approximately 1% if adjustments are made (Table 51). Rater 4 is identified given that an atypical rater was identified only 56.4% of the time for the

GHeP_rog model based on adjusted comparisons and about 60% of the time for the adjusted comparisons (Table 52).

Table 51. Overall Probability (%) of Identifying any Rater as the Atypical Rater for Data Simulated Assuming Marginal Heterogeneity when the Data were Analyzed Assuming Marginal Homogeneity

Model	**Unadjusted**	**Bonferroni**	**Holm's - Bonferroni**	**Sidak**	**Holm's - Sidak**
GHeP-rog	12.0	1.1	1.1	1.1	1.2
GHeP-Atyp4a	6.8	0.20	0.21	0.20	0.22
GHeP-Atyp4b	35.7	33.6	33.8	33.6	33.8

Table 52. Conditional Probability (%) of Identifying Rater 4 as Atypical for Data Simulated Assuming Marginal Heterogeneity when the Data were Analyzed Assuming Marginal Homogeneity

Model	**Unadjusted**	**Bonferroni**	**Holm's - Bonferroni**	**Sidak**	**Holm's - Sidak**
GHeP-rog	56.4	58.4	59.6	58.4	61.1
GHeP-Atyp4a	95.8	>99	>99	>99	100
GHeP-Atyp4b	>99	>99	>99	>99	>99

Analysis Assuming Marginal Heterogeneity. The number of times each possible pair-wise comparison was statistically significant across the 1,000 simulated contingency tables is summarized in Table 53. The majority of adjusted significant pair-wise comparisons involve Rater 4. Very few pair-wise comparisons that did not involve Rater 4 remained significant after using a multiple comparison procedure.

Table 53. Number (%) of Times Each Possible Pair-wise Comparison was Statistically Significant Across 1000 Tables Simulated under the GHeP-rog Agreement Model with Marginal Heterogeneity when Data were Analyzed Assuming Marginal Heterogeneity

Possible Pair-wise Comparisons (N)	**15 (360)**	**10 (434)**	**6 (173)**	**3 (30)**	**1 (3)**
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{2}}$					
Unadjusted	37 (10.3)	24 (7.4)	4 (6.0)	0	--
Bonferroni	1 (0.2)	2 (0.6)	1 (1.5)	0	--
Holm's- Bonferroni	1 (0.2)	2 (0.6)	2 (3.0)	0	--
Sidak	1 (0.2)	2 (0.6)	1 (1.5)	0	--
Holm's-Sidak	1 (0.2)	2 (0.6)	2 (3.0)	0	--
Missing	0	113	107	29	3
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{3}}$					
Unadjusted	26 (7.2)	9 (6.2)	1 (7.1)	0	--
Bonferroni	2 (0.6)	0	0	0	--
Holm's- Bonferroni	2 (0.6)	0	0	0	--
Sidak	2 (0.6)	0	0	0	--
Holm's-Sidak	2 (0.6)	0	0	0	--
Missing	0	289	159	28	3
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{4}}$					
Unadjusted	154 (42.7)	162 (46.4)	38 (42.2)	4 (44.4)	1 (100)
Bonferroni	62 (17.2)	83 (23.7)	27 (30.0)	3 (33.3)	1 (100)
Holm's- Bonferroni	63 (17.5)	86 (24.6)	30 (33.3)	3 (33.3)	1 (100)
Sidak	64 (17.8)	84 (24.1)	27 (30.0)	3 (33.3)	1 (100)
Holm's-Sidak	65 (18.1)	86 (24.6)	30 (33.3)	3 (33.3)	1 (100)
Missing	0	85	83	0	2
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{5}}$					
Unadjusted	24 (6.7)	11 (4.1)	2 (5.0)	0	--
Bonferroni	2 (0.6)	2 (0.8)	1 (2.5)	0	--
Holm's- Bonferroni	2 (0.6)	2 (0.8)	1 (2.5)	0	--
Sidak	2 (0.6)	2 (0.8)	1 (2.5)	0	--
Holm's-Sidak	2 (0.6)	2 (0.8)	1 (2.5)	0	--
Missing	0	165	133	30	3

Table 53 (continued)

Possible Pair-wise Comparisons (N)	**15 (360)**	**10 (434)**	**6 (173)**	**3 (30)**	**1 (3)**
$\hat{d}_5^{\overline{1}}$ vs. $\hat{d}_5^{\overline{6}}$					
Unadjusted	26 (7.2)	26 (8.0)	12 (16.7)	0	--
Bonferroni	0	0	3 (4.2)	0	--
Holm's- Bonferroni	0	1 (0.3)	3 (4.2)	0	--
Sidak	0	0	3 (4.2)	0	--
Holm's-Sidak	0	1 (0.3)	3 (4.2)	0	--
Missing	0	110	101	25	3
$\hat{d}_5^{\overline{2}}$ vs. $\hat{d}_5^{\overline{3}}$					
Unadjusted	32 (8.9)	17 (8.7)	5 (17.8)	0	--
Bonferroni	0	3 (1.6)	1 (3.6)	0	--
Holm's- Bonferroni	0	3 (1.6)	1 (3.6)	0	--
Sidak	0	3 (1.6)	1 (3.6)	0	--
Holm's-Sidak	0	3 (1.6)	1 (3.6)	0	--
Missing	0	238	145	28	3
$\hat{d}_5^{\overline{2}}$ vs. $\hat{d}_5^{\overline{4}}$					
Unadjusted	167 (46.4)	190 (47.5)	88 (65.2)	9 (69.2)	--
Raw	80 (22.2)	100 (25.0)	57 (42.2)	6 (46.2)	--
Bonferroni	83 (23.1)	100 (25.0)	59 (43.7)	6 (46.2)	--
Holm's- Bonferroni	80 (22.2)	100 (25.0)	57 (42.2)	6 (46.2)	--
Sidak	84 (23.2)	100 (25.0)	59 (43.7)	6 (46.2)	--
Holm's-Sidak	0	34	38	17	3
Missing					
$\hat{d}_5^{\overline{2}}$ vs. $\hat{d}_5^{\overline{5}}$					
Unadjusted	34 (9.4)	22 (6.8)	6 (8.0)	0	--
Bonferroni	1 (0.3)	2 (0.6)	1 (1.3)	0	--
Holm's- Bonferroni	1 (0.3)	2 (0.6)	1 (1.3)	0	--
Sidak	1 (0.3)	2 (0.6)	1 (1.3)	0	--
Holm's-Sidak	1 (0.3)	2 (0.6)	1 (1.3)	0	--
Missing	0	114	98	27	3

Table 53 (continued)

Possible Pair-wise Comparisons	**15**	**10**	**6**	**3**	**1**
(N)	**(360)**	**(434)**	**(173)**	**(30)**	**(3)**
$\hat{d}_5^{\overline{2}}$ vs. $\hat{d}_5^{\overline{6}}$					
Unadjusted	27 (7.5)	24 (6.4)	12 (10.6)	0	--
Bonferroni	1 (0.3)	2 (0.5)	5 (4.4)	0	--
Holm's- Bonferroni	1 (0.3)	3 (0.8)	5 (4.4)	0	--
Sidak	1 (0.3)	2 (0.5)	5 (4.4)	0	--
Holm's-Sidak	1 (0.3)	3 (0.8)	5 (4.4)	0	--
Missing	0	59	60	23	3
$\hat{d}_5^{\overline{3}}$ vs. $\hat{d}_5^{\overline{4}}$					
Unadjusted	114 (31.7)	69 (30.8)	18 (41.8)	3 (42.8)	--
Bonferroni	42 (11.7)	33 (14.7)	9 (20.9)	2 (28.6)	--
Holm's- Bonferroni	44 (12.2)	34 (15.2)	10 (23.3)	2 (28.6)	--
Sidak	43 (11.9)	34 (15.2)	9 (20.9)	2 (28.6)	--
Holm's-Sidak	46 (12.8)	35 (15.6)	10 (23.3)	2 (28.6)	--
Missing	0	210	130	23	3
$\hat{d}_5^{\overline{3}}$ vs. $\hat{d}_5^{\overline{5}}$					
Unadjusted	19 (5.3)	10 (6.9)	2 (14.3)	0	--
Bonferroni	2 (0.5)	2 (1.4)	0	0	--
Holm's- Bonferroni	3 (0.8)	2 (1.4)	0	0	--
Sidak	2 (0.5)	2 (1.4)	0	0	--
Holm's-Sidak	3 (0.8)	2 (1.4)	0	0	--
Missing	0	290	159	29	3
$\hat{d}_5^{\overline{3}}$ vs. $\hat{d}_5^{\overline{6}}$					
Unadjusted	24 (6.7)	18 (9.1)	2 (6.7)	0	--
Bonferroni	1 (0.3)	0	2 (6.7)	0	--
Holm's- Bonferroni	1 (0.3)	2 (1.0)	2 (6.7)	0	--
Sidak	1 (0.3)	0	2 (6.7)	0	--
Holm's-Sidak	1 (0.3)	2 (1.0)	2 (6.7)	0	--
Missing	0	235	143	28	3

Table 53 (continued)

Possible Pair-wise Comparisons (N)	**15 (360)**	**10 (434)**	**6 (173)**	**3 (30)**	**1 (3)**
$\hat{d}_5^{\overline{4}}$ vs. $\hat{d}_5^{\overline{5}}$ Unadjusted Bonferroni Holm's- Bonferroni Sidak Holm's-Sidak Missing	156 (43.3) 71 (19.7) 75 (20.8) 72 (20.0) 75 (20.8) 0	144 (41.3) 59 (17.0) 62 (17.8) 60 (17.2) 63 (18.1) 86	51 (52.0) 28 (28.6) 30 (30.6) 29 (29.6) 31 (31.6) 75	8 (72.7) 5 (45.4) 6 (54.6) 5 (45.4) 6 (54.6) 19	1 (100) 1 (100) 1 (100) 1 (100) 1 (100) 2
$\hat{d}_5^{\overline{4}}$ vs. $\hat{d}_5^{\overline{6}}$ Unadjusted Unadjusted Bonferroni Holm's- Bonferroni Sidak Holm's-Sidak Missing	184 (51.1) 96 (26.7) 97 (26.9) 96 (26.7) 98 (27.2) 0	219 (54.3) 130 (32.2) 131 (32.5) 132 (32.8) 133 (33.0) 31	88 (62.4) 57 (40.4) 59 (41.8) 57 (40.4) 59 (41.8) 32	9 (45.0) 7 (35.0) 7 (35.0) 7 (35.0) 7 (35.0) 10	1 (100) 1 (100) 1 (100) 1 (100) 1 (100) 2
$\hat{d}_5^{\overline{5}}$ vs. $\hat{d}_5^{\overline{6}}$ Unadjusted Bonferroni Holm's- Bonferroni Sidak Holm's-Sidak Missing	20 (5.6) 1 (0.3) 1 (0.3) 1 (0.3) 1 (0.3) 0	23 (7.1) 4 (1.3) 4 (1.3) 4 (1.3) 4 (1.3) 111	8 (10.1) 1 (1.3) 2 (2.5) 1 (1.3) 2 (2.5) 94	1 (16.7) 0 1 (16.7) 0 1 (16.7) 24	-- -- -- -- -- 3

The power to identify Rater 4 as the atypical rater when $d_5^{\bar{4}}$ differs from one or more other $d_5^{\bar{l}}$ is summarized in Table 54. Using a criterion that at least one comparison involving $d_5^{\bar{4}}$ has to be statistically significant, the power is 79.8% for the unadjusted comparisons compared to slightly more than 50% using the Bonferroni, Holm's- Bonferroni, Sidak or Holm's-Sidak adjustments.

Table 54. Power to Identify Rater 4 as the Atypical Rater Using Various Criteria By Multiple Comparison Procedure for GHeP-rog Scenario Simulated Assuming Marginal Heterogeneity when the Data were and Analyzed Assuming Marginal Heterogeneity

	Multiple Comparison Procedure				
Rater 4 Differs from:	**Unadjusted**	**Bonferroni**	**Holm's Bonferroni**	**Sidak**	**Holm's Sidak**
One rater	0.234	0.245	0.229	0.229	0.231
Two raters	0.225	0.159	0.163	0.163	0.165
Three raters	0.192	0.083	0.09	0.089	0.091
Four raters	0.106	0.028	0.032	0.032	0.031
Five raters	0.041	0.008	0.009	0.009	0.01
At least one rater	0.798	0.523	0.523	0.523	0.528

* 0.07% not evaluable

Table 55 summarizes the proportion of simulations that incorrectly identified the atypical rater. For the GHeP-rog model, the incorrect rater is identified in 25.6% of the simulations based on unadjusted comparisons but only 2.11% of simulations based on adjusted comparisons.

Table 55. Proportion (%) of Simulations Identifying the Incorrect Rater as Atypical for Scenarios Simulated Assuming Marginal Heterogeneity when the Data were Analyzed Assuming Marginal Heterogeneity

At least one rater is different	Unadjusted	Bonferroni	Holm's Bonferroni	Sidak	Holm's Sidak
GHeP-rog	25.6	2.11	2.11	2.11	2.11
GHeP-atyp4a	4.84	0.001	0.001	0.001	0.001
GHeP-atyp4b	3.61	< 0.001	< 0.001	< 0.001	< 0.001

The overall probability that any rater is identified as an atypical rater for the GHeP-rog model is 21.9% using unadjusted comparisons 9% if adjustments are made (Table 56). The probability that Rater 4 is correctly identified as the atypical rater given that an atypical rater was identified is 78.7% based on unadjusted comparisons but better than 95% if any of the four adjustment procedures are used (Table 57). The adjusted conditional probability provides more accurate inference than the unadjusted conditional probability.

Table 56. Overall Probability (%) of Identifying any Rater as the Atypical Rater for Data Simulated Assuming Marginal Heterogeneity when the Data were and Analyzed Assuming Marginal Heterogeneity

Model	**Unadjusted**	**Bonferroni**	**Holm's - Bonferroni**	**Sidak**	**Holm's-Sidak**
GHeP-rog	21.9	9.2	9.6	9.3	9.7
GHeP-atyp4a	15.3	5.1	5.4	5.2	5.6
GHeP-atyp4b	4.19	0.57	0.65	0.57	0.65

Table 57. Conditional Probability (%) of Identifying Rater 4 as the Atypical Rater for Data Simulated Assuming Marginal Heterogeneity when the Data were Analyzed Assuming Marginal Heterogeneity

Model	**Unadjusted**	**Bonferroni**	**Holm's-Bonferroni**	**Sidak**	**Holm's-Sidak**
GHeP-rog	78.7	95.7	95.1	95.8	95.1
GHeP-atyp4a	97.4	>99	>99	>99	>99
GHeP-atyp4b	>99	>99	>99	>99	>99

4.2.9. Simulated GHeP-atyp4a Agreement Model Assuming Marginal Heterogeneity

Generation of Simulated Tables. One thousand 2^6 contingency tables were generated under the assumption of marginal heterogeneity using the parameter estimates for the GHeP-atyp4a model shown in Table 11, column 5. The total number of counts per table ranged from 37 to 111 (Table 34). One example of the simulated cell counts of the 64 possible rating patterns for the generated 2^6 contingency tables was presented in Table 35 (col. 4). Only 16 of the 66 rating patterns in this sample simulation represented global agreement, and 30 (~ 45%) ratings patterns represented partial agreement. Eleven of the 30 partial agreement ratings represented disagreement by Rater 4 only.

The rater agreement characteristics across the 1,000 simulated contingency tables for the GHeP-atyp4a agreement model are summarized in Table 58. There is less variability in the mean marginal percentages of heterogeneous partial agreement between Raters 1, 2, 3, 5 and 6 than that seen for the GHeP-rog scenario because the parameter estimates of the $d_5^{\bar{i}}$ for i =1, 2, 3, 5 and 6 used to generate the data are constrained to be the same (2.13). The mean percentage of five-way agreement was ~3.4% when Raters 1, 2, 3, 5 or 6 was in disagreement and 9.0% when Rater 4 was in disagreement.

Table 58. Marginal Percentages for Different Category Specific Agreement Patterns and Rater Exclusion for the GHeP-atyp4a Agreement Model Simulated under the Assumption of Marginal Heterogeneity

GHeP-atyp4a Model – Heterogeneity									
Rater *i*	Marginal % for Absence	Marginal % for Presence	G %, d_6	G for Absence %, $d_{6,0}$	G for Presence %, $d_{6,1}$	GP %, d_5	GP for Absence %, $d_{5,0}$	GP for Presence %, $d_{5,1}$	Excluded Rater, % $d_5^{\ddot{i}}$
	Mean Marginal % (SE) [min,max]		**Mean Proportion (SE) [min,max]**						
1	77.1 (7.1) [50,96.2]	22.9 (7.1) [3.8,50.0]	41.5 (0.3) [17.6,66.2]	34.1 (0.2) [14.2,56.3]	7.4 (0.1) [0,27.4]	26.1 (0.2) [6.2,51.7]	18.7 (0.2) [2.4,39.7]	7.4 (0.1) [0,25.0]	3.6 (0.1) [0,13.8]
2	67.7 (8.0) [39.4,92.3]	32.3 (8.0) [7.7,60.6]							3.5 (0.1) [0,16.2]
3	71.0 (7.8) [41.4,92.1]	28.9 (7.8) [7.9,58.6]							3.5 (0.1) [0,16.9]
4	43.8 (8.2) [20.8,68.3]	56.1 (8.2) [31.7,79.2]							9.0 (0.1) [0,30.4]
5	72.4 (7.7) [42.9,94.5]	27.6 (7.7) [5.5,57.1]							3.2 (0.1) [0,11.9]
6	74.3 (7.2) [51.2,91.8]	25.7 (7.2) [8.2,48.8]							3.2 (0.1) [0,12.5]

* For agreement patterns see Table 4. G= Global agreement; GP= Global and partial agreement

Simulation scenario GHeP-atyp4a had the highest percent (61.1%) of simulated tables with no sufficient statistic for a heterogeneous partial agreement parameter equal to zero. Only seven of the simulated contingency tables had the sufficient statistic for $\boldsymbol{d}_5^{\bar{4}}$ equal to zero (Table 37).

Analysis Assuming Marginal Homogeneity. The number of times each possible pair-wise comparison was statistically significant across the 1,000 simulated contingency tables is summarized in Table 59. A majority of the significant unadjusted pair-wise comparisons and the only significant adjusted pair-wise comparisons involved Rater 4.

The power to identify Rater 4 as the atypical rater when $\boldsymbol{d}_5^{\bar{4}}$ differs from one or more of the other $\boldsymbol{d}_5^{\bar{i}}$ is summarized in Table 60. Using a criterion that requires at least one comparison involving $\boldsymbol{d}_5^{\bar{4}}$ being statistically significant, the power is 32.4% based on unadjusted comparisons and 0.9% using any of the four multiple comparison procedures considered.

Table 59. Number (%) of Times Each Possible Pair-wise Comparison was Statistically Significant Across 1000 Tables Simulated under the GHeP-Atyp4a Agreement Model Assuming Marginal Heterogeneity when the Data were Analyzed Assuming Marginal Homogeneity

Possible Pair-wise Comparisons (N)	**15 (611)**	**10 (268)**	**6 (105)**	**3 (14)**	**1 (2)**
$\hat{d}_5^1$ vs. $\hat{d}_5^2$					
Unadjusted	5 (0.8)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	93	74	11	2
$\hat{d}_5^1$ vs. $\hat{d}_5^3$					
Unadjusted	3 (0.5)	1 (0.5)	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	89	63	13	2
$\hat{d}_5^1$ vs. $\hat{d}_5^4$					
Unadjusted	106 (17.4)	35 (16.0)	18 (27.7)	0	--
Bonferroni	0	0	4 (6.2)	0	--
Holm's- Bonferroni	1 (0.1)	0	4 (6.2)	0	--
Sidak	0	0	4 (6.2)	0	--
Holm's-Sidak	1 (0.1)	0	4 (6.2)	0	--
Missing	0	50	40	4	2
$\hat{d}_5^1$ vs. $\hat{d}_5^5$					
Unadjusted	3 (0.5)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	108	69	10	2

Table 59 (continued)

Possible Pair-wise Comparisons	**15**	**10**	**6**	**3**	**1**
(N)	**(611)**	**(268)**	**(105)**	**(14)**	**(2)**
$\hat{d}_{5}^{\overline{1}}$ vs. $\hat{d}_{5}^{\overline{6}}$					
Unadjusted	3 (0.5)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	100	78	10	2
$\hat{d}_{5}^{\overline{2}}$ vs. $\hat{d}_{5}^{\overline{3}}$					
Unadjusted	3 (0.5)	1 (0.6)	0	--	--
Bonferroni	0	0	0	--	--
Holm's- Bonferroni	0	0	0	--	--
Sidak	0	0	0	--	--
Holm's-Sidak	0	0	0	--	--
Missing	0	96	73	14	2
$\hat{d}_{5}^{\overline{2}}$ vs. $\hat{d}_{5}^{\overline{4}}$					
Unadjusted	113 (18.5)	37 (17.5)	16 (28.6)	0	--
Bonferroni	2 (0.3)	0	5 (8.9)	0	--
Holm's- Bonferroni	2 (0.3)	0	5 (8.9)	0	--
Sidak	2 (0.3)	0	5 (8.9)	0	--
Holm's-Sidak	2 (0.3)	0	5 (8.9)	0	--
Missing	0	57	49	12	2
$\hat{d}_{5}^{\overline{2}}$ vs. $\hat{d}_{5}^{\overline{5}}$					
Unadjusted	5 (0.8)	2 (1.3)	0	--	--
Bonferroni	0	0	0	--	--
Holm's- Bonferroni	0	0	0	--	--
Sidak	0	0	0	--	--
Holm's-Sidak	0	0	0	--	--
Missing	0	115	75	14	2

Table 59 (continued)

Possible Pair-wise Comparisons (N)	**15 (611)**	**10 (268)**	**6 (105)**	**3 (14)**	**1 (2)**
$\hat{d}_5^{\overline{2}}$ vs. $\hat{d}_5^{\overline{6}}$					
Unadjusted	2 (0.3)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	107	83	13	2
$\hat{d}_5^{\overline{3}}$ vs. $\hat{d}_5^{\overline{4}}$					
Unadjusted	109 (17.8)	39 (18.1)	18 (24.6)	0	--
Bonferroni	2 (0.3)	0	4 (5.4)	0	--
Holm's- Bonferroni	2 (0.3)	0	4 (5.4)	0	--
Sidak	2 (0.3)	0	4 (5.4)	0	--
Holm's-Sidak	2 (0.3)	0	4 (5.4)	0	--
Missing	0	53	32	10	2
$\hat{d}_5^{\overline{3}}$ vs. $\hat{d}_5^{\overline{5}}$					
Unadjusted	1 (0.1)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	111	63	13	2
$\hat{d}_5^{\overline{3}}$ vs. $\hat{d}_5^{\overline{6}}$					
Unadjusted	3 (0.5)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	103	72	12	2

Table 59 (continued)

Possible Pair-wise Comparisons (N)	**15 (611)**	**10 (268)**	**6 (105)**	**3 (14)**	**1 (2)**
$\hat{d}_5^{\bar{4}}$ vs. $\hat{d}_5^{\bar{5}}$					
Unadjusted	113 (18.5)	33 (16.8)	16 (25.0)	0	0
Bonferroni	2 (0.3)	0	2 (3.1)	0	0
Holm's- Bonferroni	2 (0.3)	0	2 (3.1)	0	0
Sidak	2 (0.3)	0	2 (3.1)	0	0
Holm's-Sidak	2 (0.3)	1 (0.5)	2 (3.1)	0	0
Missing	0	72	41	9	2
$\hat{d}_5^{\bar{4}}$ vs. $\hat{d}_5^{\bar{6}}$					
Unadjusted	111 (18.2)	33 (16.8)	12 (23.5)	0	--
Bonferroni	1 (0.2)	1 (0.5)	2 (3.9)	0	--
Holm's- Bonferroni	1 (0.2)	1 (0.5)	2 (3.9)	0	--
Sidak	1 (0.2)	1 (0.5)	2 (3.9)	0	--
Holm's-Sidak	1 (0.2)	1 (0.5)	2 (3.9)	0	--
Missing	0	64	54	9	2
$\hat{d}_5^{\bar{5}}$ vs. $\hat{d}_5^{\bar{6}}$					
Unadjusted	4 (0.7)	0	0	--	--
Bonferroni	0	0	0	--	--
Holm's- Bonferroni	0	0	0	--	--
Sidak	0	0	0	--	--
Holm's-Sidak	0	0	0	--	--
Missing	0	122	79	14	2

Table 60. Power to Identify Rater 4 as the Atypical Rater Using Various Criteria by Multiple Comparison Procedure Simulated Assuming Marginal Heterogeneity when the Data were Analyzed Assuming Marginal Homogeneity

	Multiple Comparison Procedure				
Rater 4 Differs from:	**Unadjusted**	**Bonferroni**	**Holm's Bonferroni**	**Sidak**	**Holm's Sidak**
One rater	0.077	0.001	0.001	0.001	0
Two raters	0.064	0.001	0.001	0.001	0.002
Three raters	0.056	0.006	0.006	0.006	0.006
Four raters	0.051	0.001	0	0.001	0
Five raters	0.066	0	0.001	0	0.001
At least one rater	0.324	0.009	0.009	0.009	0.009

The overall probability that a rater other than Rater 4 is identified as an atypical rater is 2.6%, unadjusted, for the GHeP-atyp4a model, and 0 if adjusted (Table 50). The corresponding probabilities that any rater is identified are slightly higher (Table 51). The conditional probability that Rater 4 is identified as the atypical rater given that an atypical rater was identified is >99% either unadjusted or adjusted (Table 52).

Analysis Assuming Marginal Heterogeneity. The number of times each possible pair-wise comparison was statistically significant across the 1,000 simulated contingency tables is summarized in Table 61. The vast majority of unadjusted pair-wise comparisons involved Rater 4. Except for one significant $\hat{d}_5^{\overline{1}}$ vs. $\hat{d}_5^{\overline{6}}$ comparison, the only significant adjusted pair-wise comparisons involved Rater 4.

Table 61. Number (%) of Times Each Possible Pair-wise Comparison was Statistically Significant Across 1000 Tables Simulated and Analyzed under the GHeP-atyp4a Agreement Model Assuming Marginal Heterogeneity

Possible Pair-wise Comparisons (N)	**15 (611)**	**10 (268)**	**6 (105)**	**3 (14)**	**1 (2)**
$\hat{d}_5^{\overline{1}}$ vs. $\hat{d}_5^{\overline{2}}$					
Unadjusted	4 (0.7)	4 (2.3)	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	93	74	11	2
$\hat{d}_5^{\overline{1}}$ vs. $\hat{d}_5^{\overline{3}}$					
Unadjusted	3 (0.5)	0	1 (2.3)	1 (100)	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	89	63	13	2
$\hat{d}_5^{\overline{1}}$ vs. $\hat{d}_5^{\overline{4}}$					
Unadjusted	281 (46.0)	92 (42.2)	27 (41.5)	2 (20.0)	--
Bonferroni	99 (16.2)	31 (14.2)	8 (12.3)	1 (10.0)	--
Holm's- Bonferroni	105 (17.2)	34 (15.6)	8 (12.3)	1 (10.0)	--
Sidak	100 (16.4)	31 (14.2)	8 (12.3)	1 (10.0)	--
Holm's-Sidak	106 (17.4)	36 (16.5)	8 (12.3)	1 (10.0)	--
Missing	0	50	40	4	2
$\hat{d}_5^{\overline{1}}$ vs. $\hat{d}_5^{\overline{5}}$					
Unadjusted	2 (0.3)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	108	69	10	2

Table 61 (continued)

Possible Pair-wise Comparisons (N)	**15 (611)**	**10 (268)**	**6 (105)**	**3 (14)**	**1 (2)**
$\hat{d}_5^{\overline{1}}$ vs. $\hat{d}_5^{\overline{6}}$					
Unadjusted	1 (0.1)	0	1 (3.7)	0	--
Bonferroni	0	0	1 (3.7)	0	--
Holm’s- Bonferroni	0	0	1 (3.7)	0	--
Sidak	0	0	1 (3.7)	0	--
Holm’s-Sidak	0	0	1 (3.7)	0	--
Missing	0	100	78	10	2
$\hat{d}_5^{\overline{2}}$ vs. $\hat{d}_5^{\overline{3}}$					
Unadjusted	7 (1.2)	1 (0.6)	0	--	--
Bonferroni	0	0	0	--	--
Holm’s- Bonferroni	0	0	0	--	--
Sidak	0	0	0	--	--
Holm’s-Sidak	0	0	0	--	--
Missing	0	96	73	14	2
$\hat{d}_5^{\overline{2}}$ vs. $\hat{d}_5^{\overline{4}}$					
Unadjusted	259 (42.4)	82 (38.8)	21 (37.5)	0	--
Bonferroni	82 (13.4)	26 (12.3)	9 (16.1)	0	--
Holm’s- Bonferroni	90 (14.7)	29 (13.7)	9 (16.1)	0	--
Sidak	82 (13.4)	26 (12.3)	9 (16.1)	0	--
Holm’s-Sidak	90 (14.7)	29 (13.7)	9 (16.1)	0	--
Missing	0	57	49	12	2
$\hat{d}_5^{\overline{2}}$ vs. $\hat{d}_5^{\overline{5}}$					
Unadjusted	2 (0.3)	2 (1.3)	0	--	--
Bonferroni	0	0	0	--	--
Holm’s- Bonferroni	0	0	0	--	--
Sidak	0	0	0	--	--
Holm’s-Sidak	0	0	0	--	--
Missing	0	115	75	14	2

Table 61 (continued)

Possible Pair-wise Comparisons (N)	**15 (611)**	**10 (268)**	**6 (105)**	**3 (14)**	**1 (2)**
$\hat{d}_{5}^{\overline{2}}$ vs. $\hat{d}_{5}^{\overline{6}}$					
Unadjusted	4 (0.7)	1 (0.6)	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	107	83	13	2
$\hat{d}_{5}^{\overline{3}}$ vs. $\hat{d}_{5}^{\overline{4}}$					
Unadjusted	262 (42.8)	83 (38.6)	26 (35.6)	2 (50.0)	--
Bonferroni	99 (16.2)	20 (9.3)	11 (15.1)	2 (50.0)	--
Holm's- Bonferroni	103 (16.8)	24 (11.2)	13 (17.8)	2 (50.0)	--
Sidak	99 (16.2)	21 (9.8)	11 (15.1)	2 (50.0)	--
Holm's-Sidak	104 (17.0)	24 (11.2)	13 (17.8)	2 (50.0)	--
Missing	0	53	32	10	2
$\hat{d}_{5}^{\overline{3}}$ vs. $\hat{d}_{5}^{\overline{5}}$					
Unadjusted	5 (0.8)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	111	63	13	2
$\hat{d}_{5}^{\overline{3}}$ vs. $\hat{d}_{5}^{\overline{6}}$					
Unadjusted	5 (0.8)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	103	72	12	2

Table 61 (continued)

Possible Pair-wise Comparisons (N)	**15 (611)**	**10 (268)**	**6 (105)**	**3 (14)**	**1 (2)**
$\hat{d}_{5}^{\overline{4}}$ vs. $\hat{d}_{5}^{\overline{5}}$					
Unadjusted	256 (41.9)	78 (39.8)	24 (37.5)	0	0
Bonferroni	85 (13.9)	32 (16.3)	15 (23.4)	0	0
Holm's- Bonferroni	87 (14.2)	33 (16.8)	15 (23.4)	0	0
Sidak	86 (14.1)	33 (16.8)	15 (23.4)	0	0
Holm's-Sidak	90 (14.7)	33 (16.8)	15 (23.4)	0	0
Missing	0	72	41	9	2
$\hat{d}_{5}^{\overline{4}}$ vs. $\hat{d}_{5}^{\overline{6}}$					
Unadjusted	273 (44.6)	78 (38.2)	17 (33.3)	3 (60.0)	--
Bonferroni	82 (13.4)	30 (14.7)	8 (15.7)	2 (40.0)	--
Holm's- Bonferroni	88 (14.4)	30 (14.7)	8 (15.7)	2 (40.0)	--
Sidak	85 (13.9)	30 (14.7)	8 (15.7)	2 (40.0)	--
Holm's-Sidak	91 (14.9)	30 (14.7)	9 (17.7)	2 (40.0)	--
Missing	0	64	54	9	2
$\hat{d}_{5}^{\overline{5}}$ vs. $\hat{d}_{5}^{\overline{6}}$					
Unadjusted	5 (0.8)	0	0	--	--
Bonferroni	0	0	0	--	--
Holm's- Bonferroni	0	0	0	--	--
Sidak	0	0	0	--	--
Holm's-Sidak	0	0	0	--	--
Missing	0	129	79	14	2

The power to identify Rater 4 as the atypical rater when $d_5^{\bar{4}}$ differs from one or more other $d_5^{\bar{i}}$ is summarized in Table 62. Using a criterion that requires at least one comparison involving $d_5^{\bar{4}}$ to be statistically significant, the power to identify Rater 4 as the atypical rater is reduced from 68.8% (unadjusted) and ~ 32% when a multiple comparison procedure is used. Approximately 5% of the unadjusted pair-wise comparisons and 0.1% of the adjusted pair-wise comparisons identified the incorrect rater (Table 55).

Table 62. Power to Identify Rater 4 as the Atypical Rater Using Various Criteria by Multiple Comparison Procedure for the GHeP-atyp4a Scenario Simulated Assuming Marginal Heterogeneity when the Data were Analyzed Assuming Marginal Heterogeneity

	Multiple Comparison Procedure				
Rater 4 Differs from:	**Unadjusted**	**Bonferroni**	**Holm's Bonferroni**	**Sidak**	**Holm's Sidak**
One rater	0.184	0.134	0.125	0.136	0.124
Two raters	0.143	0.088	0.081	0.089	0.085
Three raters	0.132	0.052	0.059	0.053	0.059
Four raters	0.13	0.028	0.032	0.028	0.033
Five raters	0.099	0.014	0.019	0.014	0.019
At least one rater	0.688	0.31.6	0.316	0.32	0.32

The overall probability that any rater is identified as an atypical rater is 15.3% if unadjusted pair-wise comparisons are used and ~ 5% if adjusted pair-wise comparisons are used (Table 56). The probability that Rater 4 is the atypical rater given an atypical rater was identified is > 99% with or without adjustment for multiple comparisons (Table 57).

4.2.10. Simulated GHeP-atyp4b Agreement Model Assuming Marginal Heterogeneity

Generation of Simulated Tables. One thousand 2^6 contingency tables were generated under the assumption of marginal heterogeneity using the parameter estimates for the GHeP-atyp4b agreement model shown in Table 11, column 6. The total number of counts per table ranged from 57 to 224 (Table 34); the mode of 110 is approximately 1.6 times the sample size of intestinal biopsy data. One example of the simulated cell counts of the 64 possible rating patterns for the generated 2^6 contingency tables was presented in Table 35 (col. 5). Only 32 of the 121 rating patterns in this one simulation represented global agreement, whereas 63 (~ 52%) ratings patterns represented partial agreement. Thirty-two of the 63 partial agreement ratings were because Rater 4 was in disagreement.

The rater agreement characteristics across the 1,000 simulated contingency tables for the GHeP-atyp4b model are summarized in Table 63. The marginal percentages for the absence of mucosecretion diminution were ~81% for Raters 1, 2, 3, 5, and 6 and 30.7% for Rater 4. The mean percentage of global agreement was only 28.8%, representing predominantly global agreement on the absence of the lesion (23.8%). The partial agreement of 49.1% also represented predominately partial agreement on the absence of the lesion (43.9%). The mean marginal percentage of five-way agreement when Rater 1, 2, 3, 5, or 6 is excluded is ~2.3%, and 37.3% when Rater 4 is excluded. These highly skewed percentages are because of the parameter estimates used for the simulation scenario.

Table 63. Marginal Percentages for Different Category Specific Agreement Patterns and Rater Exclusion for the GHeP-atyp4b Agreement Model Simulated under the Assumption of Marginal Heterogeneity

GHeP-atyp4b Model – Heterogeneity										
Rater *i*	Marginal % for Absence	Marginal % for Presence	G %, d_6	G for Absence %, $d_{6,0}$	G for Presence %, $d_{6,1}$	GP %, d_5	GP for Absence %, $d_{5,0}$	GP for Presence %, $d_{5,1}$	Excluded Rater, % $d_5^{\ddot{i}}$
	Mean Marginal % (SE) [min,max]		**Mean Proportion (SE) [min,max]**						
1	84.3 (5.4) [65.4,95.4]	15.6 (5.4) [4.5,34.6]	28.8 (0.2) [11.3,53.2]	23.8 (0.2) [7.6,46.7]	5.0 (0.1) [0,17.3]	49.1 (0.3) [21.7,80.4]	43.9 (0.3) [18.6,75.9]	2.5 (0.1) [0,11.5]	2.5 (0.1) [0,11.5]
2	78.0 (6.3) [52.9,93.2]	22.0 (6.3) [6.7,47.1]							2.4 (0.1) [0,10.1]
3	80.2 (6.1) [56.7,93.7]	19.8 (6.1) [6.3,43.3]							2.2 (0.1) [0,10.3]
4	30.7 (7.6) [11.3,57.8]	69.2 (7.6) [42.1,88.6]							37.3 (0.3) [12.0,73.2]
5	81.1 (6.0) [59.0,96.0]	18.9 (6.0) [4.0,40.9]							2.3 (0.1) [0,9.2]
6	82.0 (5.7) [61.1,97.2]	18.0 (5.7) [2.8,38.9]							2.3 (0.1) [0,11.5]

* For agreement patterns see Table 4. G= Global agreement; GP= Global and partial agreement

Under simulation scenario GHeP-atyp4b, 61.1% of simulated tables had no sufficient statistic for a heterogeneous partial agreement parameter equal to zero. None of the simulated contingency tables had the sufficient statistic for $d_5^{\bar{4}}$ equal to zero (Table 35).

Analysis Assuming Marginal Homogeneity. The number of times each possible pair-wise comparison was statistically significant across the 1,000 simulated contingency tables is summarized in Table 64. Almost all unadjusted pair-wise comparisons involving Rater 4 were statistically significant, as were most of the adjusted comparisons. The dramatically higher percentage of significant adjusted pair-wise comparisons in the GHeP-atyp4b vs. the GHeP-atyp4a simulation scenario is the result of increasing the parameter estimate $\hat{d}_5^{\bar{4}}$ from 0.36 to 2.21.

Table 64. Number (%) of Times Each Possible Pair-wise Comparison was Statistically Significant Across 1000 Tables Simulated under the GHeP-atyp4b Agreement Model with Marginal Heterogeneity and Data were Analyzed Assuming Marginal Homogeneity

Possible Pair-wise Comparisons (N)	**15 (609)**	**10 (283)**	**6 (88)**	**3 (17)**	**1 (3)**
$\hat{d}_{5}^{\bar{1}}$ vs. $\hat{d}_{5}^{\bar{2}}$					
Unadjusted	5 (0.8)	0	0	1 (33.3)	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	105	56	14	3
$\hat{d}_{5}^{\bar{1}}$ vs. $\hat{d}_{5}^{\bar{3}}$					
Unadjusted	5 (0.8)	1 (0.6)	0	--	--
Bonferroni	0	0	0	--	--
Holm's- Bonferroni	0	0	0	--	--
Sidak	0	0	0	--	--
Holm's-Sidak	0	0	0	--	--
Missing	0	113	64	17	3
$\hat{d}_{5}^{\bar{1}}$ vs. $\hat{d}_{5}^{\bar{4}}$					
Unadjusted	605 (99.3)	235 (99.1)	54 (98.2)	6 (100)	1 (100)
Bonferroni	567 (93.1)	224 (94.5)	54 (98.2)	6 (100)	1 (100)
Holm's- Bonferroni	569 (93.4)	225 (94.9)	54 (98.2)	6 (100)	1 (100)
Sidak	567 (93.1)	224 (94.5)	54 (98.2)	6 (100)	1 (100)
Holm's-Sidak	569 (93.4)	225 (94.9)	54 (98.2)	6 (100)	1 (100)
Missing	0	46	33	11	2
$\hat{d}_{5}^{\bar{1}}$ vs. $\hat{d}_{5}^{\bar{5}}$					
Unadjusted	1 (0.1)	1 (0.6)	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	105	59	14	3

Table 64 (continued)

Possible Pair-wise Comparisons (N)	**15 (609)**	**10 (283)**	**6 (88)**	**3 (17)**	**1 (3)**
$\hat{d}_5^{\bar{1}}$ vs. $\hat{d}_5^{\bar{6}}$					
Unadjusted	1 (0.1)	0	0	--	--
Bonferroni	0	0	0	--	--
Holm's- Bonferroni	0	0	0	--	--
Sidak	0	0	0	--	--
Holm's-Sidak	0	0	0	--	--
Missing	0	98	63	17	3
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{3}}$					
Unadjusted	1 (0.1)	1 (0.6)	0	--	--
Bonferroni	0	0	0	--	--
Holm's- Bonferroni	0	0	0	--	--
Sidak	0	0	0	--	--
Holm's-Sidak	0	0	0	--	--
Missing	0	126	60	17	3
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{4}}$					
Unadjusted	605 (99.3)	223 (99.6)	57 (100)	8 (100)	--
Bonferroni	576 (94.5)	216 (96.4)	57 (100)	7 (87.5)	--
Holm's- Bonferroni	578 (94.9)	217 (96.8)	57 (100)	7 (87.5)	--
Sidak	576 (94.5)	216 (96.4)	57 (100)	7 (87.5)	--
Holm's-Sidak	578 (94.9)	217 (96.8)	57 (100)	7 (87.5)	--
Missing	0	59	31	9	3
$\hat{d}_5^{\bar{2}}$ vs. $\hat{d}_5^{\bar{5}}$					
Unadjusted	4 (0.6)	0	0	--	--
Bonferroni	0	0	0	--	--
Holm's- Bonferroni	0	0	0	--	--
Sidak	0	0	0	--	--
Holm's-Sidak	0	0	0	--	--
Missing	0	118	59	17	3

Table 64 (continued)

Possible Pair-wise Comparisons (N)	**15 (609)**	**10 (283)**	**6 (88)**	**3 (17)**	**1 (3)**
$\hat{d}_{5}^{\overline{2}}$ vs. $\hat{d}_{5}^{\overline{6}}$					
Unadjusted	5 (0.8)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	107	63	14	3
$\hat{d}_{5}^{\overline{3}}$ vs. $\hat{d}_{5}^{\overline{4}}$					
Unadjusted	608 (99.8)	214 (99.1)	48 (97.9)	5 (100)	--
Bonferroni	568 (93.3)	208 (96.3)	47 (95.9)	4 (80.0)	--
Holm's- Bonferroni	575 (94.4)	209 (96.7)	47 (95.9)	4 (80.0)	--
Sidak	568 (93.3)	208 (96.3)	47 (95.9)	4 (80.0)	--
Holm's-Sidak	575 (94.4)	209 (96.7)	47 (95.9)	4 (80.0)	--
Missing	0	67	39	12	3
$\hat{d}_{5}^{\overline{3}}$ vs. $\hat{d}_{5}^{\overline{5}}$					
Unadjusted	1 (0.1)	1 (0.6)	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	126	64	15	3
$\hat{d}_{5}^{\overline{3}}$ vs. $\hat{d}_{5}^{\overline{6}}$					
Unadjusted	1 (0.1)	1 (0.6)	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	119	66	16	3

Table 64 (continued)

Possible Pair-wise Comparisons (N)	**15 (609)**	**10 (283)**	**6 (88)**	**3 (17)**	**1 (3)**
$\hat{d}_{5}^{\overline{4}}$ vs. $\hat{d}_{5}^{\overline{5}}$					
Unadjusted	606 (99.5)	224 (100)	53 (98.2)	8 (100)	1 (100)
Bonferroni	560 (92.0)	212 (94.6)	52 (96.3)	8 (100)	1 (100)
Holm's- Bonferroni	568 (93.3)	212 (94.6)	52 (96.3)	8 (100)	1 (100)
Sidak	561 (92.1)	212 (94.6)	52 (96.3)	8 (100)	1 (100)
Holm's-Sidak	568 (93.3)	212 (94.6)	52 (96.3)	8 (100)	1 (100)
Missing	0	59	34	9	2
$\hat{d}_{5}^{\overline{4}}$ vs. $\hat{d}_{5}^{\overline{6}}$					
Unadjusted	605 (99.3)	230 (99.6)	49 (100)	7 (100)	1 (100)
Bonferroni	570 (93.6)	219 (94.8)	48 (98.0)	7 (100)	1 (100)
Holm's- Bonferroni	573 (94.1)	220 (95.2)	48 (98.0)	7 (100)	1 (100)
Sidak	570 (93.6)	219 (94.8)	48 (98.0)	7 (100)	1 (100)
Holm's-Sidak	573 (94.1)	220 (95.2)	48 (98.0)	7 (100)	1 (100)
Missing	0	52	39	10	2
$\hat{d}_{5}^{\overline{5}}$ vs. $\hat{d}_{5}^{\overline{6}}$					
Unadjusted	3 (0.5)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm's- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm's-Sidak	0	0	0	0	--
Missing	0	111	62	14	3

Power to identify Rater 4 as the atypical rater when $d_5^{\bar{4}}$ differs from one or more of the other $d_5^{\bar{i}}$ is summarized in Table 65. Using the criteria that at least one of the pair-wise comparisons of the heterogeneous partial agreement parameters is significant, both the unadjusted and adjusted approaches provide better than 96% power to identify Rater 4 as atypical. Only 2.3% of unadjusted comparisons and no adjusted comparisons identified the incorrect rater as atypical (Table 50).

The overall probability that any rater is identified as atypical rater is ~34% whether or not adjustments for the number of comparisons are made (Table 51). The probability that Rater 4 is the atypical rater given that an atypical rater was identified is >99% either unadjusted or adjusted for multiple comparisons (Table 52).

Table 65. Power to Identify Rater 4 as the Atypical Rater Using Various Criteria By Multiple Comparison Procedure for the GHeP-atyp4b Scenario Simulated Assuming Marginal Heterogeneity when the Data were Analyzed Assuming Marginal Homogeneity

	Multiple Comparison Procedure				
Rater 4 Differs from:	**Unadjusted**	**Bonferroni**	**Holm's Bonferroni**	**Sidak**	**Holm's Sidak**
One rater	0.004	0.01	0.01	0.01	0.01
Two raters	0.018	0.024	0.02	0.024	0.02
Three raters	0.089	0.105	0.096	0.105	0.096
Four raters	0.288	0.275	0.279	0.274	0.279
Five raters	0.599	0.548	0.557	0.549	0.557
At least one rater	0.998	0.962	0.962	0.962	0.962

Analysis Assuming Marginal Heterogeneity. The number of times each possible pair-wise comparison was statistically significant across the 1,000 simulated contingency tables is summarized in Table 66. In contrast to Table 64, relatively few unadjusted or adjusted pair-wise comparisons involving Rater 4 were statistically significant.

Table 66. Number (%) of Times Each Possible Pair-wise Comparison was Statistically Significant Across 1000 Tables Simulated and Analyzed under the GHeP-atyp4b Agreement Model Assuming Marginal Heterogeneity

Possible Pair-wise Comparisons (N)	**15 (609)**	**10 (283)**	**6 (88)**	**3 (17)**	**1 (3)**
$\hat{d}_5^1$ vs. $\hat{d}_5^2$					
Unadjusted	4 (0.6)	0	0	0	--
Bonferroni	0	0	0	0	--
Holm’s- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm’s-Sidak	0	0	0	0	--
Missing	0	105	56	14	3
$\hat{d}_5^1$ vs. $\hat{d}_5^3$					
Unadjusted	6 (1.0)	1 (0.6)	0	--	--
Bonferroni	0	0	0	--	--
Holm’s- Bonferroni	0	0	0	--	--
Sidak	0	0	0	--	--
Holm’s-Sidak	0	0	0	--	--
Missing	0	113	64	17	3
$\hat{d}_5^1$ vs. $\hat{d}_5^4$					
Unadjusted	70 (11.5)	27 (11.4)	2 (3.6)	1 (16.7)	0
Bonferroni	11 (1.8)	3 (1.3)	1 (1.8)	1 (16.7)	0
Holm’s- Bonferroni	11 (1.8)	3 (1.3)	1 (1.8)	1 (16.7)	0
Sidak	11 (1.8)	3 (1.3)	1 (1.8)	1 (16.7)	0
Holm’s-Sidak	11 (1.8)	3 (1.3)	1 (1.8)	1 (16.7)	0
Missing	0	46	33	11	2
$\hat{d}_5^1$ vs. $\hat{d}_5^5$					
Unadjusted	3 (0.5)	2 (1.1)	0	0	--
Bonferroni	0	0	0	0	--
Holm’s- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm’s-Sidak	0	0	0	0	--
Missing	0	105	59	14	3

Table 66 (continued)

Possible Pair-wise Comparisons (N)	**15 (609)**	**10 (283)**	**6 (88)**	**3 (17)**	**1 (3)**
$\hat{d}_5^{\overline{1}}$ vs. $\hat{d}_5^{\overline{6}}$					
Unadjusted	8 (1.3)	0	0	--	--
Bonferroni	0	0	0	--	--
Holm's- Bonferroni	0	0	0	--	--
Sidak	0	0	0	--	--
Holm's-Sidak	0	0	0	--	--
Missing	0	98	63	17	3
$\hat{d}_5^{\overline{2}}$ vs. $\hat{d}_5^{\overline{3}}$					
Unadjusted	3 (0.5)	0	0	--	--
Bonferroni	0	0	0	--	--
Holm's- Bonferroni	0	0	0	--	--
Sidak	0	0	0	--	--
Holm's-Sidak	0	0	0	--	--
Missing	0	126	60	17	3
$\hat{d}_5^{\overline{2}}$ vs. $\hat{d}_5^{\overline{4}}$					
Unadjusted	56 (9.2)	26 (11.6)	5 (8.8)	0	--
Bonferroni	10 (1.6)	2 (0.9)	2 (3.5)	0	--
Holm's- Bonferroni	11 (1.8)	4 (1.8)	2 (3.5)	0	--
Sidak	10 (1.6)	2 (0.9)	2 (3.5)	0	--
Holm's-Sidak	11 (1.8)	4 (1.8)	2 (3.5)	0	--
Missing	0	59	31	9	3
$\hat{d}_5^{\overline{2}}$ vs. $\hat{d}_5^{\overline{5}}$					
Unadjusted	7 (1.2)	1 (0.6)	0	--	--
Bonferroni	0	0	0	--	--
Holm's- Bonferroni	0	0	0	--	--
Sidak	0	0	0	--	--
Holm's-Sidak	0	0	0	--	--
Missing	0	118	59	17	3

Table 66 (continued)

Possible Pair-wise Comparisons (N)	**15 (609)**	**10 (283)**	**6 (88)**	**3 (17)**	**1 (3)**
$\hat{d}_{5}^{\overline{2}}$ vs. $\hat{d}_{5}^{\overline{6}}$					
Unadjusted	3 (0.5)	1 (0.6)	0	0	--
Bonferroni	0	0	0	0	--
Holm’s- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm’s-Sidak	0	0	0	0	--
Missing	0	107	63	14	3
$\hat{d}_{5}^{\overline{3}}$ vs. $\hat{d}_{5}^{\overline{4}}$					
Unadjusted	64 (10.5)	26 (12.0)	5 (10.2)	1 (20.0)	--
Bonferroni	11 (1.8)	2 (0.9)	2 (4.0)	1 (20.0)	--
Holm’s- Bonferroni	13 (2.1)	2 (0.9)	2 (4.0)	1 (20.0)	--
Sidak	11 (1.8)	2 (0.9)	2 (4.0)	1 (20.0)	--
Holm’s-Sidak	13 (2.1)	2 (0.9)	2 (4.0)	1 (20.0)	--
Missing	0	67	39	12	3
$\hat{d}_{5}^{\overline{3}}$ vs. $\hat{d}_{5}^{\overline{5}}$					
Unadjusted	2 (0.3)	1 (0.6)	1 (4.2)	0	--
Bonferroni	0	0	0	0	--
Holm’s- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm’s-Sidak	0	0	0	0	--
Missing	0	126	64	15	3
$\hat{d}_{5}^{\overline{3}}$ vs. $\hat{d}_{5}^{\overline{6}}$					
Unadjusted	3 (0.5)	1 (0.6)	0	0	--
Bonferroni	0	0	0	0	--
Holm’s- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm’s-Sidak	0	0	0	0	--
Missing	0	119	66	16	3

Table 66 (continued)

Possible Pair-wise Comparisons (N)	**15 (609)**	**10 (283)**	**6 (88)**	**3 (17)**	**1 (3)**
$\hat{d}_{5}^{\overline{4}}$ vs. $\hat{d}_{5}^{\overline{5}}$					
Unadjusted	64 (10.5)	24 (10.7)	3 (5.5)	1 (12.5)	0
Bonferroni	11 (1.8)	2 (0.9)	1 (1.9)	1 (12.5)	0
Holm’s- Bonferroni	11 (1.8)	3 (1.3)	1 (1.9)	1 (12.5)	0
Sidak	11 (1.8)	2 (0.9)	1 (1.9)	1 (12.5)	0
Holm’s-Sidak	11 (1.8)	3 (1.3)	1 (1.9)	1 (12.5)	0
Missing	0	59	34	9	2
$\hat{d}_{5}^{\overline{4}}$ vs. $\hat{d}_{5}^{\overline{6}}$					
Unadjusted	77 (12.6)	25 (10.8)	0	0	0
Bonferroni	8 (1.3)	3 (1.3)	0	0	0
Holm’s- Bonferroni	10 (1.6)	4 (1.7)	0	0	0
Sidak	8 (1.3)	3 (1.3)	0	0	0
Holm’s-Sidak	10 (1.6)	4 (1.7)	0	0	0
Missing	0	52	39	10	2
$\hat{d}_{5}^{\overline{5}}$ vs. $\hat{d}_{5}^{\overline{6}}$					
Unadjusted	1 (0.1)	1 (0.6)	0	0	--
Bonferroni	0	0	0	0	--
Holm’s- Bonferroni	0	0	0	0	--
Sidak	0	0	0	0	--
Holm’s-Sidak	0	0	0	0	--
Missing	0	111	62	14	3

The power to identify Rater 4 as the atypical rater when $d_5^{\bar{4}}$ differs from one or more of the other $d_5^{\bar{i}}$ is summarized in Table 67. The power is low to detect Rater 4 based on unadjusted comparisons (20.0%) or adjusted comparisons (3.3%) based on a criterion that at least one of the five pair-wise comparisons of the heterogeneous partial agreement parameters is significant. Using unadjusted pair-wise comparisons, 3.61% of the simulations identify an incorrect rater while none of the adjusted pair-wise comparisons identify a rater other than Rater 4 as the atypical rater (Table 55). The overall probability that *any* rater is identified as an atypical rater is 4.19% if unadjusted pair-wise comparisons are used and less than 1% if adjusted pair-wise comparisons are used (Table 56). The probability that Rater 4 is identified as the atypical rater given that an atypical rater was identified is >99% using either unadjusted or adjusted pair-wise comparisons (Table 57).

Table 67. Power to Identify Rater 4 as the Atypical Rater Using Various Criteria By Multiple Comparison Procedure for the GHeP-atyp4b Scenario Simulated Assuming Marginal Heterogeneity when the Data were Analyzed Assuming Marginal Heterogeneity

Rater 4 Differs from:	Unadjusted	Bonferroni	Holm's Bonferroni	Sidak	Holm's Sidak
One rater	0.08	0.015	0.012	0.015	0.012
Two raters	0.041	0.007	0.007	0.007	0.007
Three raters	0.025	0.004	0.004	0.004	0.004
Four raters	0.030	0.004	0.007	0.004	0.007
Five raters	0.024	0.003	0.033	0.033	0.033
At least one rater	0.20	0.033	0.033	0.033	0.033

4.3. SUMMARY

Because the results for each simulated scenario were comparable for the four multiple comparison procedures considered, the results for only the unadjusted and Holm's-Bonferroni procedures are summarized in Tables 68 through Table 73. Tables 68-70, respectively, summarize the probabilities of identifying Rater 4 and identifying a rater other than 4, and the conditional probability of identifying Rater 4 given that at least one rater was identified as atypical, all for data simulated assuming marginal homogeneity. For the G and GP scenarios simulated under the assumption of marginal homogeneity, the Type I error is virtually zero to detect either Rater 4 (Table 68) or any rater other than Rater 4 (Table 69) for both the unadjusted and Holm's-Bonferroni adjusted comparisons. Using unadjusted pair-wise comparisons, the power to identify the Rater 4 correctly as atypical rater was about 27% for the GHeP-rog and GHeP-atyp4a models and increased to 44.2% when the Rater 4 effect was exaggerated (Table 68). Very few of the unadjusted comparisons in Table 69 identified a rater other than Rater 4 as atypical for either the GHeP-rog or GHeP-atyp4b models; however, 17.1% of the simulations incorrectly identified an atypical rater for the GHeP-atyp4a model. The power was extremely low (less than 3%) for each of the corresponding Holms-Bonferroni adjusted comparisons in Tables 68 and 69. For both the unadjusted and adjusted pair-wise comparisons, the corresponding conditional power to identify Rater 4 correctly as atypical is high (>94%) for the GHeP-rog and GHeP-atyp4b models, but less than 61% for the GHeP-atyp4a model (Table 70).

Table 68. Proportion (%) of Simulations That Identify Rater 4 as the Atypical Rater for Scenarios Simulated Assuming Marginal Homogeneity

	Analyzed Assuming Marginal Homogeneity	
Model	**Unadjusted**	**Holm's – Bonferroni**
G	<0.1	<0.1
GP	<0.1	<0.1
GHeP-rog	27.7	0.6
GHeP-atyp4a	27.5	2.3
GHeP-atyp4b	44.2	2.9

Table 69. Proportion (%) of Simulations That Identify a Rater Other Than Rater 4 as the Atypical Rater for Scenarios Simulated Assuming Marginal Homogeneity

	Analyzed Assuming Marginal Homogeneity	
Model	**Unadjusted**	**Holm's – Bonferroni**
G	<0.1	<0.1
GP	<0.1	<0.1
GHeP-rog	0.7	<0.1
GHeP-atyp4a	17.1	1.9
GHeP-atyp4b	2.6	<0.1

Table 70. Conditional Probability (%) of Identifying Rater 4 as Atypical Given That At Least One Rater Was Identified for Scenarios Simulated Assuming Marginal Homogeneity

	Analyzed Assuming Marginal Homogeneity	
Model	**Unadjusted**	**Holm's – Bonferroni**
G	na	na
GP	na	na
GHeP-rog	94.4	> 99
GHeP-atyp4a	60.7	55.5
GHeP-atyp4b	97.1	> 99

na= not applicable

Tables 71-73, respectively, summarize the probabilities of identifying Rater 4 and identifying a rater other than 4, and the conditional probability of identifying Rater 4 given that at least one rater was identified as atypical, all for data simulated assuming marginal heterogeneity. Two different GHeP models were fit to each set of simulated data: one model which incorrectly assumed marginal homogeneity and another which correctly assumed marginal heterogeneity. Under the assumption of marginal homogeneity, the Type I error of the unadjusted pair-wise comparisons is 6.6% for the G model and 58.8% for the GP model (Table 71). This 58.8% appears to be picking up the marginal heterogeneity due to Rater 4 in the partial agreement parameters, because the marginal heterogeneity is ignored in the fitted model. The power of the unadjusted comparisons to detect Rater 4 is about 35% for the GHeP-rog and GHep-atyp4a scenarios; the power of both the unadjusted and Holms-Bonferroni comparisons is >95% for the GHeP-atyp4b scenario. The Holms-Bonferroni procedure is quite conservative for the other scenarios analyzed assuming marginal homogeneity.

When these simulated data are analyzed assuming marginal heterogeneity, the Type I error of the unadjusted comparisons is approximately twice the nominal level for the G and GP models; the Holm's Bonferroni procedure is somewhat conservative for these models (Table 71). The power to detect Rater 4 as atypical using unadjusted comparisons in the GHeP-rog and GHeP-atyp4a models is improved considerably when the correct analytic model is assumed. The power for the GHeP-atyp4b model is unexpectedly low.

For the GP model analyzed assuming marginal heterogeneity, the probability of identifying the wrong rater is 5.1% using unadjusted comparisons (Table 72). Among the GHeP models, only for the GHeP-rog model with unadjusted comparisons does the probability of detecting the wrong rater exceed the nominal level.

Except for the GHeP-rog model analyzed assuming marginal homogeneity, the conditional power was very high to correctly identify Rater 4 when at least one rater was identified as atypical (Table 73). In this GHeP-rog model, the Holm's Bonferroni procedure actually had somewhat higher conditional power than the unadjusted comparisons assuming both marginal homogeneity and marginal heterogeneity.

Table 71. Proportion (%) of Simulations That Identify Rater 4 as the Atypical Rater for Scenarios Simulated Assuming Marginal Heterogeneity

	Analysis Assuming Marginal Homogeneity		**Analysis Assuming Marginal Heterogeneity**	
Model	**Unadjusted**	**Holm's –Bonferroni**	**Unadjusted**	**Holm's –Bonferroni**
G	6.6	0.7	11.0	2.9
GP	58.8	4.3	9.6	1.5
GHeP-rog	35.2	3.2	79.8	52.3
GHeP-atyp4a	32.4	0.9	68.8	31.6
GHeP-atyp4b	99.8	96.2	20.0	3.3

Table 72. Proportion (%) of Simulations That Identify a Rater Other Than Rater 4 as the Atypical Rater for Scenarios Simulated Assuming Marginal Heterogeneity

	Analysis Assuming Marginal Homogeneity		**Analysis Assuming Marginal Heterogeneity**	
Model	**Unadjusted**	**Holm's –Bonferroni**	**Unadjusted**	**Holm's –Bonferroni**
G	<0.1	<0.1	<0.1	<0.1
GP	<0.1	<0.1	5.1	<0.1
GHeP-rog	25.7	2.7	25.6	2.1
GHeP-atyp4a	2.6	<0.1	4.8	<0.1
GHeP-atyp4b	2.3	<0.1	3.6	<0.1

Table 73. Conditional Probability (%) of Identifying Rater 4 as Atypical Given That At Least One Rater Was Identified for Scenarios Simulated Assuming Marginal Heterogeneity

	Analysis Assuming Marginal Homogeneity		**Analysis Assuming Marginal Heterogeneity**	
Model	**Unadjusted**	**Holm's –Bonferroni**	**Unadjusted**	**Holm's –Bonferroni**
G	> 99	> 99	> 99	> 99
GP	> 99	> 99	82.7	> 99
GHeP-rog	56.4	59.6	78.7	95.1
GHeP-atyp4a	95.8	> 99	97.4	> 99
GHeP-atyp4b	> 99	> 99	> 99	> 99

5. DISCUSSION

Rogel et al. (1998) proposed using the heterogeneous partial agreement parameters in a log-linear model to address the problem of identifying an atypical rater in the context of a best-fitting model. Their work focused on model selections issues, and did not provide specific guidance with respect to identifying particular raters. The present work formalizes inferential procedures to identify an atypical rater using pair-wise comparisons of the heterogeneous partial agreement parameters, with particular attention paid to the issue of multiple comparisons due to the relatively large number of possible pair-wise comparisons. The Type I error and power of the proposed procedures are assessed in a simulation study, assuming either marginal homogeneity or marginal heterogeneity across raters. In the models considered, agreement was aggregated across categories of the outcome so that the approach is not sensitive to the prevalence of the outcome.

This study provides no evidence of elevated Type I error for unadjusted pair-wise comparisons of the heterogeneous partial agreement parameters assuming marginal homogeneity. While the unconditional power to identify the designated atypical rater is low for data simulated assuming marginal homogeneity, the conditional power is high using either unadjusted or adjusted comparisons for the unconstrained scenario and the scenario with the effect of the atypical rater exaggerated.

This study provides evidence that the use of unadjusted pair-wise comparisons of the heterogeneous partial agreement parameters is anti-conservative and the use of adjusted pair-wise comparisons is conservative assuming either marginal homogeneity or heterogeneity for the global model simulated under the assumption of marginal heterogeneity. The GP model is interesting because the heterogeneous partial agreement parameters in the incorrect analytic model (i.e. analysis assuming marginal homogeneity) appear to correctly identify Rater 4 for the wrong reason; even though there is no true differential five-way agreement in the simulated data, Rater 4 has a different marginal distribution that is not being parameterized directly in the analysis. However, for the GHeP-rog and GHeP-atyp4a models, the power is even higher if the pair-wise comparisons are conducted within the framework of the correct (i.e. marginal heterogeneity) analytic model. At issue is whether one is overadjusting for the ways in which Rater 4 could be atypical; Rater 4 could disagree relatively more often because his/her marginal distribution is different, or could disagree and share the same marginal distribution. In these simulations the power was highest when the differences in the marginal distributions were taken into account; Rater 4 contributed relatively little to the five-way disagreement in this situation. Another strategy would be to examine differences in the marginal heterogeneity parameters in the GP or GHeP models. This was not addressed in the present work, but will be a focus of future efforts.

The simulation study was designed so that Rater 4 was the atypical rater. In a real life application, the atypical rater is not known *a priori*. Although the overall power of the proposed approach was low in many settings considered, the conditional power to correctly identify the atypical rater (given that someone was identified) was generally quite high. In some settings the identity of the atypical rater is obvious (e.g. a single rater is involved in multiple significant pair-wise comparisons). However, if two raters differ only from each other and none of the other raters differ from each other, then both raters might be considered atypical. Moreover, if an investigator has concerns about poor inter-rater agreement, corrective action can be taken in the absence of definitive statistical evidence identifying an atypical rater. Every effort should be made to improve the consistency of ratings prior to conducting the primary study that assesses the impact of an intervention.

The descriptive summary table (e.g., marginal percentages for different category specific agreement patterns) provides clinicians with a tool to help them identify an atypical rater. The clinician can determine whether or not the magnitude of the differences in these proportions attributable to each rater is of clinical concern. Confidence intervals for the heterogeneous partial agreement parameters can also aid the clinician.

There are some limitations of this research. First, it is based on only one example and the underlying structure of the data was not clearly GHeP. Second, the relatively small number of discrepant ratings limited our power to detect atypical raters and possibly the clinical importance of detecting such discrepancies. However, this is frequently the case when experienced raters are involved in a study. Although it may have limited our inferences, the number of specimens rated (68) is not unusual for inter-rater agreement studies. In this example, although almost 25% of the 5-way agreement was due to a discrepant rating by Rater 4, this corresponds to only 6 ratings. Lastly, the GHeP model considered only assesses K-1 partial agreement and ignored other kinds of disagreement. However, given sufficient data, other types of disagreement could be addressed by redefining the agreement parameters. Future work includes (i) investigating the marginal heterogeneity parameters as an alternative strategy to identify atypical raters under this scenario and (ii) generalizing the programs to account for imbalanced and/or multi-category nominal data.

In conclusion, the heterogeneous agreement parameters generally do highlight the most atypical rater in the marginal homogeneity scenarios considered, although the power is low to detect such a rater as statistically significantly different from the other raters. Inference is less straightforward in the case of marginal heterogeneity, as the marginal heterogeneity parameters may be over-controlling for the disagreement by allowing a different marginal distribution. In either case, for the scenarios considered, pair-wise comparisons of the heterogeneous partial agreement parameters are quite likely to identify the correct rater as atypical when any rater is identified.

Appendix A
Parameterization of the Rater Effect Variables Using Sum-to-Zero Constraints

```
Parameterization.log
rpattern r1 r2 r3 r4 r5 r6
```

R1 through R6 are the variables representing the main effects of each rater, Rater 1 through 6, respectively.

Rating Pattern	R1	R2	R3	R4	R5	R6
000000	-1	-1	-1	-1	-1	-1
000001	-1	-1	-1	-1	-1	1
000010	-1	-1	-1	-1	1	-1
000011	-1	-1	-1	-1	1	1
000100	-1	-1	-1	1	-1	-1
000101	-1	-1	-1	1	-1	1
000110	-1	-1	-1	1	1	-1
000111	-1	-1	-1	1	1	1
001000	-1	-1	1	-1	-1	-1
001001	-1	-1	1	-1	-1	1
001010	-1	-1	1	-1	1	-1
001011	-1	-1	1	-1	1	1
001100	-1	-1	1	1	-1	-1
001101	-1	-1	1	1	-1	1
001110	-1	-1	1	1	1	-1
001111	-1	-1	1	1	1	1
010000	-1	1	-1	-1	-1	-1
010001	-1	1	-1	-1	-1	1
010010	-1	1	-1	-1	1	-1
010011	-1	1	-1	-1	1	1
010100	-1	1	-1	1	-1	-1
010101	-1	1	-1	1	-1	1
010110	-1	1	-1	1	1	-1
010111	-1	1	-1	1	1	1
011000	-1	1	1	-1	-1	-1
011001	-1	1	1	-1	-1	1
011010	-1	1	1	-1	1	-1

Rating Pattern	R1	R2	R3	R4	R5	R6
011011	-1	1	1	-1	1	1
011100	-1	1	1	1	-1	-1
011101	-1	1	1	1	-1	1
011110	-1	1	1	1	1	-1
011111	-1	1	1	1	1	1
100000	1	-1	-1	-1	-1	-1
100001	1	-1	-1	-1	-1	1
100010	1	-1	-1	-1	1	-1
100011	1	-1	-1	-1	1	1
100100	1	-1	-1	1	-1	-1
100101	1	-1	-1	1	-1	1
100110	1	-1	-1	1	1	-1
100111	1	-1	-1	1	1	1
101000	1	-1	1	-1	-1	-1
101001	1	-1	1	-1	-1	1
101010	1	-1	1	-1	1	-1
101011	1	-1	1	-1	1	1
101100	1	-1	1	1	-1	-1
101101	1	-1	1	1	-1	1
101110	1	-1	1	1	1	-1
101111	1	-1	1	1	1	1
110000	1	1	-1	-1	-1	-1
110001	1	1	-1	-1	-1	1
110010	1	1	-1	-1	1	-1
110011	1	1	-1	-1	1	1
110100	1	1	-1	1	-1	-1
110101	1	1	-1	1	-1	1
110110	1	1	-1	1	1	-1
110111	1	1	-1	1	1	1
111000	1	1	1	-1	-1	-1
111001	1	1	1	-1	-1	1
111010	1	1	1	-1	1	-1
111011	1	1	1	-1	1	1

Rating Pattern	R1	R2	R3	R4	R5	R6
111100	1	1	1	1	-1	-1
111101	1	1	1	1	-1	1
111110	1	1	1	1	1	-1
111111	1	1	1	1	1	1

Appendix B
Parameterization of the Indicator Variables Used For the G, GP, GHeP Models

```
. list  rpattern  e6  e5  e5m1 e5m2 e5m3 e5m4 e5m5 e5m6 e5sub

     +----------------------------------------------------------------------+
     | rpattern   e6   e5   e5m1   e5m2   e5m3   e5m4   e5m5   e5m6   e5sub |
     |----------------------------------------------------------------------|
  1. |   000000    1    0      0      0      0      0      0      0       0 |
  2. |   000001    0    1      0      0      0      0      0      1       1 |
  3. |   000010    0    1      0      0      0      0      1      0       1 |
  4. |   000011    0    0      0      0      0      0      0      0       0 |
  5. |   000100    0    1      0      0      0      1      0      0       0 |
     |----------------------------------------------------------------------|
  6. |   000101    0    0      0      0      0      0      0      0       0 |
  7. |   000110    0    0      0      0      0      0      0      0       0 |
  8. |   000111    0    0      0      0      0      0      0      0       0 |
  9. |   001000    0    1      0      0      1      0      0      0       1 |
 10. |   001001    0    0      0      0      0      0      0      0       0 |
     |----------------------------------------------------------------------|
 11. |   001010    0    0      0      0      0      0      0      0       0 |
 12. |   001011    0    0      0      0      0      0      0      0       0 |
 13. |   001100    0    0      0      0      0      0      0      0       0 |
 14. |   001101    0    0      0      0      0      0      0      0       0 |
 15. |   001110    0    0      0      0      0      0      0      0       0 |
     |----------------------------------------------------------------------|
 16. |   001111    0    0      0      0      0      0      0      0       0 |
 17. |   010000    0    1      0      1      0      0      0      0       1 |
 18. |   010001    0    0      0      0      0      0      0      0       0 |
 19. |   010010    0    0      0      0      0      0      0      0       0 |
 20. |   010011    0    0      0      0      0      0      0      0       0 |
     |----------------------------------------------------------------------|
 21. |   010100    0    0      0      0      0      0      0      0       0 |
 22. |   010101    0    0      0      0      0      0      0      0       0 |
 23. |   010110    0    0      0      0      0      0      0      0       0 |
 24. |   010111    0    0      0      0      0      0      0      0       0 |
 25. |   011000    0    0      0      0      0      0      0      0       0 |
     |----------------------------------------------------------------------|
 26. |   011001    0    0      0      0      0      0      0      0       0 |
 27. |   011010    0    0      0      0      0      0      0      0       0 |
 28. |   011011    0    0      0      0      0      0      0      0       0 |
 29. |   011100    0    0      0      0      0      0      0      0       0 |
 30. |   011101    0    0      0      0      0      0      0      0       0 |
     |----------------------------------------------------------------------|
 31. |   011110    0    0      0      0      0      0      0      0       0 |
 32. |   011111    0    1      1      0      0      0      0      0       1 |
 33. |   100000    0    1      1      0      0      0      0      0       1 |
 34. |   100001    0    0      0      0      0      0      0      0       0 |
 35. |   100010    0    0      0      0      0      0      0      0       0 |
     |----------------------------------------------------------------------|
 36. |   100011    0    0      0      0      0      0      0      0       0 |
 37. |   100100    0    0      0      0      0      0      0      0       0 |
 38. |   100101    0    0      0      0      0      0      0      0       0 |
 39. |   100110    0    0      0      0      0      0      0      0       0 |
 40. |   100111    0    0      0      0      0      0      0      0       0 |
     |----------------------------------------------------------------------|
 41. |   101000    0    0      0      0      0      0      0      0       0 |
 42. |   101001    0    0      0      0      0      0      0      0       0 |
 43. |   101010    0    0      0      0      0      0      0      0       0 |
 44. |   101011    0    0      0      0      0      0      0      0       0 |
 45. |   101100    0    0      0      0      0      0      0      0       0 |
     |----------------------------------------------------------------------|
 46. |   101101    0    0      0      0      0      0      0      0       0 |
 47. |   101110    0    0      0      0      0      0      0      0       0 |
 48. |   101111    0    1      0      1      0      0      0      0       1 |
 49. |   110000    0    0      0      0      0      0      0      0       0 |
 50. |   110001    0    0      0      0      0      0      0      0       0 |
     |----------------------------------------------------------------------|
     +----------------------------------------------------------------------+
     | rpattern   e6   e5   e5m1   e5m2   e5m3   e5m4   e5m5   e5m6   e5sub |
     |----------------------------------------------------------------------|
 51. |   110010    0    0      0      0      0      0      0      0       0 |
 52. |   110011    0    0      0      0      0      0      0      0       0 |
 53. |   110100    0    0      0      0      0      0      0      0       0 |
 54. |   110101    0    0      0      0      0      0      0      0       0 |
 55. |   110110    0    0      0      0      0      0      0      0       0 |
     |----------------------------------------------------------------------|
 56. |   110111    0    1      0      0      1      0      0      0       1 |
 57. |   111000    0    0      0      0      0      0      0      0       0 |
 58. |   111001    0    0      0      0      0      0      0      0       0 |
 59. |   111010    0    0      0      0      0      0      0      0       0 |
 60. |   111011    0    1      0      0      0      1      0      0       0 |
     |----------------------------------------------------------------------|
 61. |   111100    0    0      0      0      0      0      0      0       0 |
 62. |   111101    0    1      0      0      0      0      1      0       1 |
 63. |   111110    0    1      0      0      0      0      0      1       1 |
 64. |   111111    1    0      0      0      0      0      0      0       0 |
     +----------------------------------------------------------------------+
```

Appendix C
Parameter Estimates and and Variance-Covariance Matrices for the G, GP, GHeP Models Under the Assumption of Marginal Homogeneity & Heterogeneity

* GLOBAL AGREEMENT MODEL

```
. glm ctdm e6 , f(p)
note: ctdm has non-integer values

Iteration 0:   log likelihood = -100.61721
Iteration 1:   log likelihood = -90.578469
Iteration 2:   log likelihood =  -90.53458
Iteration 3:   log likelihood = -90.534554
Iteration 4:   log likelihood = -90.534554

Generalized linear models                          No. of obs      =        64
Optimization     : ML: Newton-Raphson              Residual df     =        62
                                                   Scale parameter =         1
Deviance         =  120.2993083                    (1/df) Deviance =  1.940311
Pearson          =  134.9754386                    (1/df) Pearson  =  2.177023

Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]
Standard errors  : OIM

Log likelihood   = -90.53455362                    AIC             =  2.891705
BIC              = -137.5514428

------------------------------------------------------------------------------
        ctdm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          e6 |   3.197598   .2442317    13.09   0.000     2.718913    3.676284
       _cons |  -.4895482   .1622214    -3.02   0.003    -.8074964   -.1716001
------------------------------------------------------------------------------
(Variance-Covariance Matrix)
. matrix list e(V)

symmetric e(V)[2,2]
                  ctdm:       ctdm:
                    e6       _cons
   ctdm:e6   .05964912
ctdm:_cons  -.02631579   .02631579
```

* GLOBAL & PARTIAL AGREEMENT MODEL

```
. glm ctdm e6  e5 , f(p)
note: ctdm has non-integer values

Iteration 0:   log likelihood = -94.510217
Iteration 1:   log likelihood = -84.276409
Iteration 2:   log likelihood = -84.228052
Iteration 3:   log likelihood = -84.228019
Iteration 4:   log likelihood = -84.228019

Generalized linear models                          No. of obs      =        64
Optimization     : ML: Newton-Raphson              Residual df     =        61
                                                   Scale parameter =         1
Deviance         =  107.6862395                    (1/df) Deviance =  1.765348
Pearson          =  110.2901961                    (1/df) Pearson  =  1.808036

Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]
Standard errors  : OIM

Log likelihood   = -84.22801918                    AIC             =  2.725876
BIC              = -146.0056286

------------------------------------------------------------------------------
        ctdm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          e6 |   3.575551   .2845213    12.57   0.000     3.017899    4.133202
          e5 |   1.215807   .3262554     3.73   0.000     .5763585    1.855256
       _cons |  -.8675006   .2182179    -3.98   0.000      -1.2952   -.4398014
------------------------------------------------------------------------------
(Variance-Covariance Matrix)
```

```
. matrix list e(V)

symmetric e(V)[3,3]
                  ctdm:       ctdm:       ctdm:
                    e6          e5       _cons
   ctdm:e6   .08095238
   ctdm:e5   .04761905   .10644258
ctdm:_cons  -.04761905  -.04761905   .04761905
```

* GHeP-rog MODEL

```
. glm ctdm e6 e5m1 e5m2 e5m3 e5m4 e5m5 e5m6, f(p)
note: ctdm has non-integer values

Iteration 0:   log likelihood = -92.361679
Iteration 1:   log likelihood =  -81.85987
Iteration 2:   log likelihood = -81.817945
Iteration 3:   log likelihood = -81.817916
Iteration 4:   log likelihood = -81.817916

Generalized linear models                          No. of obs      =        64
Optimization     : ML: Newton-Raphson              Residual df     =        56
                                                   Scale parameter =         1
Deviance         =  102.8660327                    (1/df) Deviance =  1.836893
Pearson          =  100.1333333                    (1/df) Pearson  =  1.788095

Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]
Standard errors  : OIM

Log likelihood   = -81.81791578                    AIC             =   2.80681
BIC              =   -130.03142

------------------------------------------------------------------------------
        ctdm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          e6 |   3.575551   .2845213    12.57   0.000     3.017899    4.133202
        e5m1 |   .8675006   .7400129     1.17   0.241     -.582898    2.317899
        e5m2 |   1.272966   .6172134     2.06   0.039     .0632496    2.482682
        e5m3 |   .1743534   1.023533     0.17   0.865    -1.831734     2.18044
        e5m4 |   1.966113     .46291     4.25   0.000     1.058826      2.8734
        e5m5 |   .8675006   .7400129     1.17   0.241     -.582898    2.317899
        e5m6 |   1.272966   .6172134     2.06   0.039     .0632496    2.482682
       _cons |  -.8675006   .2182179    -3.98   0.000      -1.2952   -.4398014
------------------------------------------------------------------------------
```

```
(Variance-covariance Matrix)
. matrix list e(V)

symmetric e(V)[8,8]
                  ctdm:       ctdm:       ctdm:       ctdm:       ctdm:       ctdm:
                    e6        e5m1        e5m2        e5m3        e5m4        e5m5
   ctdm:e6   .08095238
 ctdm:e5m1   .04761905   .54761905
 ctdm:e5m2   .04761905   .04761905   .38095238
 ctdm:e5m3   .04761905   .04761905   .04761905    1.047619
 ctdm:e5m4   .04761905   .04761905   .04761905   .04761905   .21428571
 ctdm:e5m5   .04761905   .04761905   .04761905   .04761905   .04761905   .54761905
 ctdm:e5m6   .04761905   .04761905   .04761905   .04761905   .04761905   .04761905
ctdm:_cons  -.04761905  -.04761905  -.04761905  -.04761905  -.04761905  -.04761905

                  ctdm:       ctdm:
                  e5m6       _cons
 ctdm:e5m6   .38095238
ctdm:_cons  -.04761905   .04761905
```

* GHeP-rme MODEL

```
. glm ctdm e6 e5m4 e5sub, f(p)
note: ctdm has non-integer values

Iteration 0:   log likelihood = -93.043343
Iteration 1:   log likelihood = -82.550895
Iteration 2:   log likelihood = -82.509176
Iteration 3:   log likelihood = -82.509147
Iteration 4:   log likelihood = -82.509147

Generalized linear models                          No. of obs      =        64
Optimization     : ML: Newton-Raphson              Residual df     =        60
                                                   Scale parameter =         1
Deviance         =  104.2484956                    (1/df) Deviance =  1.737475
Pearson          =  100.7393939                    (1/df) Pearson  =   1.67899

Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]
Standard errors  : OIM

Log likelihood   = -82.50914727                    AIC             =  2.703411
BIC              = -145.2844894

------------------------------------------------------------------------------
        ctdm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          e6 |   3.575551   .2845213    12.57   0.000     3.017899    4.133202
        e5m4 |   1.966113     .46291     4.25   0.000     1.058826      2.8734
       e5sub |   .9628107   .3721937     2.59   0.010     .2333245    1.692297
       _cons |  -.8675006   .2182179    -3.98   0.000      -1.2952   -.4398014
------------------------------------------------------------------------------

(Variance-covariance Matrix)

. matrix list e(V)

symmetric e(V)[4,4]
                  ctdm:       ctdm:       ctdm:       ctdm:
                    e6        e5m4       e5sub       _cons
   ctdm:e6   .08095238
 ctdm:e5m4   .04761905   .21428571
ctdm:e5sub   .04761905   .04761905   .13852814
ctdm:_cons  -.04761905  -.04761905  -.04761905   .04761905
```

Marginal Heterogeneity Models

```
. *Global (G) Model
. glm ctdm r1-r6 e6, f(p)
note: ctdm has non-integer values

Iteration 0:   log likelihood = -74.117064
Iteration 1:   log likelihood = -63.330689
Iteration 2:   log likelihood = -62.989964
Iteration 3:   log likelihood = -62.989091
Iteration 4:   log likelihood = -62.989091

Generalized linear models                          No. of obs      =        64
Optimization     : ML: Newton-Raphson              Residual df     =        56
                                                   Scale parameter =         1
Deviance         =  65.20838239                    (1/df) Deviance =  1.164435
Pearson          =  81.21747696                    (1/df) Pearson  =  1.450312

Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]
Standard errors  : OIM

Log likelihood   = -62.98909065                    AIC             =  2.218409
BIC              = -167.6890703

------------------------------------------------------------------------------
        ctdm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          r1 |  -.5051301   .1692812    -2.98   0.003    -.8369152   -.1733451
          r2 |  -.1592249   .1537164    -1.04   0.300    -.4605036    .1420538
          r3 |  -.2667977   .1567927    -1.70   0.089    -.5741059    .0405104
          r4 |   .8001227   .1984658     4.03   0.000     .4111369    1.189109
          r5 |   -.322779   .1590265    -2.03   0.042    -.6344652   -.0110927
          r6 |  -.3807889   .1617964    -2.35   0.019     -.697904   -.0636738
          e6 |   3.474279    .343287    10.12   0.000     2.801449    4.147109
       _cons |  -1.080286    .254489    -4.24   0.000    -1.579075   -.5814965
------------------------------------------------------------------------------

. matrix list e(V)

symmetric e(V)[8,8]
                  ctdm:       ctdm:       ctdm:       ctdm:       ctdm:       ctdm:
                    r1          r2          r3          r4          r5          r6
   ctdm:r1   .02865612
   ctdm:r2  -.00340322   .02362874
   ctdm:r3  -.00353864  -.00284647   .02458397
   ctdm:r4  -.00622486  -.00466073  -.00499298   .03938867
   ctdm:r5  -.00364282  -.00294219  -.00307166  -.00521623   .02528944
   ctdm:r6  -.00377685  -.00306288  -.00319335   -.0054877  -.00329183   .02617806
   ctdm:e6  -.00820788   .00216985  -.00078424   .04452793  -.00238668  -.00412201
ctdm:_cons   .01371901   .00241488   .00552442  -.03578212   .00725241   .00915215

                  ctdm:       ctdm:
                    e6       _cons
   ctdm:e6   .11784594
ctdm:_cons  -.06320704   .06476465

. *Global &Partial Agreement (GP) Model
. glm ctdm r1-r6 e6 e5, f(p)
note: ctdm has non-integer values

Iteration 0:   log likelihood = -66.964091
Iteration 1:   log likelihood = -56.744469
Iteration 2:   log likelihood = -56.570797
Iteration 3:   log likelihood = -56.570167
Iteration 4:   log likelihood = -56.570167

Generalized linear models                          No. of obs      =        64
Optimization     : ML: Newton-Raphson              Residual df     =        55
                                                   Scale parameter =         1
```

```
Deviance         =   52.3705352                    (1/df) Deviance =  .9521915
Pearson          =  54.70780162                    (1/df) Pearson  =  .9946873

Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]
Standard errors  : OIM

Log likelihood   = -56.57016705                    AIC             =  2.049068
BIC              = -176.3680344

------------------------------------------------------------------------------
        ctdm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          r1 |  -.5268156   .1855831    -2.84   0.005    -.8905518   -.1630794
          r2 |  -.1040925   .1666673    -0.62   0.532    -.4307544    .2225695
          r3 |  -.2352723   .1702461    -1.38   0.167    -.5689485    .0984039
          r4 |   .8436832   .1871705     4.51   0.000     .4768358    1.210531
          r5 |  -.3038033   .1729446    -1.76   0.079    -.6427685    .0351619
          r6 |  -.3748939    .176343    -2.13   0.034    -.7205199   -.0292679
          e6 |   3.964699   .3693411    10.73   0.000     3.240804    4.688594
          e5 |   1.253173   .3339669     3.75   0.000     .5986095    1.907736
       _cons |  -1.484641   .2923541    -5.08   0.000    -2.057645   -.9116379
------------------------------------------------------------------------------

. matrix list e(V)

symmetric e(V)[9,9]
                  ctdm:       ctdm:       ctdm:       ctdm:       ctdm:       ctdm:
                    r1          r2          r3          r4          r5          r6
   ctdm:r1   .03444109
   ctdm:r2  -.00537786   .02777799
   ctdm:r3  -.00535131  -.00452496   .02898373
   ctdm:r4  -.00655518  -.00295152  -.00396841   .03503279
   ctdm:r5  -.00536261  -.00470427  -.00485101  -.00452297   .02990984
   ctdm:r6  -.00538555  -.00490789   -.0050021  -.00512808  -.00507816   .03109686
   ctdm:e6  -.01289128   .00637501   .00064844   .04057046  -.00241462  -.00566665
   ctdm:e5  -.00356208   .00624207    .0031854   .00456246   .00155263  -.00013648
ctdm:_cons   .01676927   -.0031609   .00255022  -.03336546   .00567674   .00905237

                  ctdm:       ctdm:       ctdm:
                    e6          e5       _cons
   ctdm:e6   .13641284
   ctdm:e5   .05903352   .11153389
ctdm:_cons  -.08697025  -.05186641   .08547089

. *GHeP-rog Model
. glm ctdm r1-r6 e6  e5m1 e5m2 e5m3 e5m4 e5m5 e5m6, f(p)
note: ctdm has non-integer values

Iteration 0:   log likelihood = -67.821025
Iteration 1:   log likelihood = -54.067189
Iteration 2:   log likelihood = -53.682628
Iteration 3:   log likelihood =  -53.68041
Iteration 4:   log likelihood =  -53.68041

Generalized linear models                          No. of obs      =        64
Optimization     : ML: Newton-Raphson              Residual df     =        50
                                                   Scale parameter =         1
Deviance         =  46.59102072                    (1/df) Deviance =  .9318204
Pearson          =  54.65787957                    (1/df) Pearson  =  1.093158

Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]
Standard errors  : OIM

Log likelihood   = -53.68040981                    AIC             =  2.115013
BIC              = -161.3531334

------------------------------------------------------------------------------
        ctdm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
          r1 |  -.6495583   .2034988    -3.19   0.001    -1.048409   -.2507079
          r2 |  -.2479767    .174465    -1.42   0.155    -.5899217    .0939683
          r3 |  -.3511232   .1930984    -1.82   0.069    -.7295892    .0273427
          r4 |   1.354155   .3694166     3.67   0.000     .6301121    2.078199
          r5 |  -.4240635   .1854973    -2.29   0.022    -.7876315   -.0604956
          r6 |  -.4821315   .1800043    -2.68   0.007    -.8349336   -.1293295
          e6 |   4.501199   .5460147     8.24   0.000      3.43103    5.571368
        e5m1 |   1.964981   .8503748     2.31   0.021     .2982774    3.631685
        e5m2 |   2.444098   .7949606     3.07   0.002     .8860038    4.002192
        e5m3 |    1.38638   1.130114     1.23   0.220    -.8286017    3.601362
        e5m4 |   .3662216   .5645338     0.65   0.517    -.7402442    1.472687
        e5m5 |   2.083242    .871234     2.39   0.017     .3756544    3.790829
        e5m6 |   2.476514   .7636371     3.24   0.001     .9798123    3.973215
       _cons |  -2.084366   .5139002    -4.06   0.000    -3.091592    -1.07714
------------------------------------------------------------------------------

. matrix list e(V)

symmetric e(V)[14,14]
                  ctdm:       ctdm:       ctdm:       ctdm:       ctdm:       ctdm:
                    r1          r2          r3          r4          r5          r6
   ctdm:r1   .04141178
   ctdm:r2   .00027207   .03043802
   ctdm:r3  -.00105385   .00085813     .037287
   ctdm:r4  -.03122223  -.02420736  -.03051538   .13646866
   ctdm:r5  -.00052005   .00122311   .00008172  -.02691087   .03440924
   ctdm:r6  -.00009219   .00153284   .00030505  -.02420731   .00081971   .03240157
   ctdm:e6  -.04383912  -.02037006  -.03180308   .16852089  -.02952998  -.02745081
 ctdm:e5m1    -.0155656  -.03150372  -.04060955   .14041093  -.04025279  -.03964364
 ctdm:e5m2  -.04724359  -.04209883  -.03487818   .18305387  -.03361015  -.03232511
 ctdm:e5m3  -.04861312  -.02626755  -.04306428   .17484678  -.03470091   -.0336038
 ctdm:e5m4    .0214294   .03133485   .03144249  -.10412641   .02723382   .02446856
 ctdm:e5m5  -.05014925  -.02745548  -.03675151   .16639326  -.03274803  -.03503207
 ctdm:e5m6  -.05113894  -.02823508  -.03745916    .1611632  -.03678816  -.02583819
ctdm:_cons    .0496831   .02709194   .03642927  -.16891589   .03557817   .03459983

                  ctdm:       ctdm:       ctdm:       ctdm:       ctdm:       ctdm:
                    e6        e5m1        e5m2        e5m3        e5m4        e5m5
   ctdm:e6   .29813202
 ctdm:e5m1   .20691671   .72313725
 ctdm:e5m2   .27111711    .2355624    .6319624
 ctdm:e5m3   .26224735   .23027009   .28198441   1.2771569
 ctdm:e5m4    -.066457  -.08246962  -.09670597  -.09062649   .31869836
 ctdm:e5m5   .25094658   .22243742   .27298662   .26707495  -.08649916   .75904872
 ctdm:e5m6   .24306341   .21651519   .26650602   .26077656  -.08479318   .25260006
ctdm:_cons  -.25448157  -.22496545  -.27583585  -.26982373   .08757428   -.2614067

                  ctdm:       ctdm:
                  e5m6       _cons
 ctdm:e5m6   .58314168
ctdm:_cons  -.25521999   .26409344

. *GHeP-rme Model
. glm ctdm r1-r6 e6 e5_at e5sub, f(p)
note: ctdm has non-integer values

Iteration 0:   log likelihood = -68.024168
Iteration 1:   log likelihood = -54.747051
Iteration 2:   log likelihood = -54.359522
Iteration 3:   log likelihood = -54.357354
Iteration 4:   log likelihood = -54.357354

Generalized linear models                          No. of obs      =        64
Optimization     : ML: Newton-Raphson              Residual df     =        54
                                                   Scale parameter =         1
Deviance         =  47.94490854                    (1/df) Deviance =  .8878687
Pearson          =  55.38098786                    (1/df) Pearson  =  1.025574

Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]
Standard errors  : OIM
```

```
Log likelihood   = -54.35735372                    AIC             =  2.011167
BIC              =  -176.634778

------------------------------------------------------------------------------
        ctdm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          r1 |  -.6376812   .1942569    -3.28   0.001    -1.018418   -.2569446
          r2 |  -.2351533   .1784588    -1.32   0.188    -.5849261    .1146194
          r3 |  -.3594418   .1816393    -1.98   0.048    -.7154483   -.0034352
          r4 |   1.345573   .3633264     3.70   0.000     .6334668     2.05768
          r5 |  -.4249581   .1840168    -2.31   0.021    -.7856245   -.0642918
          r6 |  -.4930534    .186926    -2.64   0.008    -.8594217   -.1266851
          e6 |   4.489447   .5368271     8.36   0.000     3.437285    5.541609
       e5_at |   .3702664    .564022     0.66   0.512    -.7351965    1.475729
       e5sub |   2.133214   .5808956     3.67   0.000     .9946795    3.271748
       _cons |  -2.075287   .5066553    -4.10   0.000    -3.068314   -1.082261
------------------------------------------------------------------------------

. matrix list e(V)

symmetric e(V)[10,10]
                  ctdm:        ctdm:        ctdm:        ctdm:        ctdm:        ctdm:
                    r1           r2           r3           r4           r5           r6
   ctdm:r1   .03773576
   ctdm:r2   .00016595    .03184754
   ctdm:r3   .00006682    .00124605    .03299285
   ctdm:r4  -.02804216   -.02572997   -.02634692    .13200608
   ctdm:r5   .00004263    .00098163    .00064065   -.02675271    .03386219
   ctdm:r6   .00004187    .00071369    .00044457   -.02719949    .00032302    .03494134
   ctdm:e6  -.03825617   -.02170036   -.02666347    .16180863    -.0294101   -.03230681
ctdm:e5_at    .0210437    .03272307    .02893164   -.10240308    .02701279    .02506357
ctdm:e5sub  -.03840928   -.03248767   -.03439781     .1598975   -.03551129   -.03663898
ctdm:_cons   .04493006    .02785026    .03269279   -.16318541    .03547502    .03848346

                  ctdm:        ctdm:        ctdm:        ctdm:
                    e6        e5_at        e5sub        _cons
   ctdm:e6   .28818333
ctdm:e5_at  -.06370685    .31812086
ctdm:e5sub   .23887007   -.08615438    .33743965
ctdm:_cons  -.24586888    .08528791   -.25056839    .25669963
```

Appendix D
SAS Macro %MVN Used to Generate Multivariate Normal Data

%MVN macro used to simulate data from a multivariate normal distribution (SAS Institute Inc; http://ftp.sas.com/techsup/download/stat/mvn.html)

```
/************************************************************************
           %MVN macro:  Generating multivariate normal data

    DISCLAIMER:
      THIS INFORMATION IS PROVIDED BY SAS INSTITUTE INC. AS A SERVICE TO
      ITS USERS.  IT IS PROVIDED "AS IS".  THERE ARE NO WARRANTIES,
      EXPRESSED OR IMPLIED, AS TO MERCHANTABILITY OR FITNESS FOR A
      PARTICULAR PURPOSE REGARDING THE ACCURACY OF THE MATERIALS OR CODE
      CONTAINED HEREIN.

    PURPOSE:
      The %MVN macro generates multivariate normal data using the
      Cholesky root of the variance-covariance matrix.  Bivariate normal
      data can be generated using the DATA step code that follows the
      macro.

    REQUIRES:
      The %MVN macro requires Version 6.06 or later of SAS/IML software.
      The DATA step code for generating bivariate normal data requires
      only Version 6.06 Base SAS software.

    USAGE:
      The macro input/output paramters are:

           VARCOV= SAS data set that contains the variance-covariance
                   (and only the variance covariance) matrix.  The macro
                   expects m variables and m observations in the data
                   set, where m is the number of variables to generate.

           MEANS=  SAS data set that contains the mean vector.  The
                   macro expects a single variable with m observations
                   containing the m means for the variables generated.

           N=      Number of observations to generate.

           SEED=   Starting seed value for the random number generator.
                   Default value is 0, which will use the system clock
                   to generate a seed.

           SAMPLE= SAS data set name for the resulting multivariate
                   normal data.  The variable names will be COL1-COLm.

    LIMITATIONS:
      No error checking is done.  The macro assumes that dataset
      names entered are valid, and exist in the case of the VARCOV=
      and MEANS= options.
    EXAMPLE:
      This example generates 1000 observations from a 3 variable
      multivariate normal distribution with specified mean vector and
      covariance matrix.

        * Store the variance-covariance matrix in a data set;
        data varcov;
           input m1-m3;
           cards;
          4 1.8   4
        1.8   9 3.6
          4 3.6  16
        ;

        * Store the mean vector in a data set ;
        data means;
           input m1;
           cards;
        10
        20
        30
        ;

        %mvn(varcov=varcov,
             means=means,
             n=1000,
             sample=test)

        proc corr data=test noprob cov;
          run;


************************************************************************/

%macro mvn(varcov=,       /* dataset for variance-covariance matrix */
           means=,        /* dataset for mean vector */
           n=,            /* sample size */
           seed=0,        /* seed for random number generator */
           sample=);      /* output dataset name */


 /* Get initial seed value.  If seed<=0, then generate seed from the
    system clock. */

data _null_;
   if &seed le 0 then do;
      seed = int(time());   /* get clock time in integer seconds */
      put seed=;
      call symput('seed',seed);   /* store seed as macro variable */
   end;
run;
```

```
 /* Generate the multivariate normal data in SAS/IML */

proc iml worksize=100;
   use &varcov;            /* read variance-covariance matrix */
   read all into cov;
   use &means;             /* read means */
   read all into mu;
   v=nrow(cov);            /* calculate number of variables */
   n=&n;
   seed = &seed;
   l=t(root(cov));         /* calculate cholesky root of cov matrix */
   z=normal(j(v,&n,&seed));/* generate nvars*samplesize normals */
   x=l*z;                  /* premultiply by cholesky root */
   x=repeat(mu,1,&n)+x;    /* add in the means */
   tx=t(x);
   create &sample from tx;  /* write out sample data to sas dataset */
   append from tx;
quit;

%mend mvn;
```

Appendix E
SAS Program Used to Generate Count Data for 2^6 Contingency Table for the GHeP Model

```
libname get "C:\aaPhDSimulations\HomoG\GHePRME";
libname g "C:\aaPhDSimulations\HomoG\GHePRME\STATAds";
/*
Dataset cvarpats contains the 64 possible rating patterns.
The covariate patterns of the 64 possible rating patterns are
enumerated.
Variables r1 through r6 represent the ratings or raters 1 thourgh 6,
respectively.
*/
data g.suffstat;
input what$;
cards;
suffstat
;
run;
data g.cvarpats;
input rp r1 r2 r3 r4 r5 r6;
cards;
1  0 0 0 0 0 0
2  0 0 0 0 0 1
3  0 0 0 0 1 0
4  0 0 0 0 1 1
5  0 0 0 1 0 0
6  0 0 0 1 0 1
7  0 0 0 1 1 0
8  0 0 0 1 1 1
9  0 0 1 0 0 0
10 0 0 1 0 0 1
11 0 0 1 0 1 0
12 0 0 1 0 1 1
13 0 0 1 1 0 0
14 0 0 1 1 0 1
15 0 0 1 1 1 0
16 0 0 1 1 1 1
17 0 1 0 0 0 0
18 0 1 0 0 0 1
19 0 1 0 0 1 0
20 0 1 0 0 1 1
21 0 1 0 1 0 0
22 0 1 0 1 0 1
23 0 1 0 1 1 0
24 0 1 0 1 1 1
25 0 1 1 0 0 0
26 0 1 1 0 0 1
27 0 1 1 0 1 0
28 0 1 1 0 1 1
29 0 1 1 1 0 0
30 0 1 1 1 0 1
31 0 1 1 1 1 0
32 0 1 1 1 1 1
33 1 0 0 0 0 0
34 1 0 0 0 0 1
35 1 0 0 0 1 0
36 1 0 0 0 1 1
37 1 0 0 1 0 0
```

```
38 1 0 0 1 0 1
39 1 0 0 1 1 0
40 1 0 0 1 1 1
41 1 0 1 0 0 0
42 1 0 1 0 0 1
43 1 0 1 0 1 0
44 1 0 1 0 1 1
45 1 0 1 1 0 0
46 1 0 1 1 0 1
47 1 0 1 1 1 0
48 1 0 1 1 1 1
49 1 1 0 0 0 0
50 1 1 0 0 0 1
51 1 1 0 0 1 0
52 1 1 0 0 1 1
53 1 1 0 1 0 0
54 1 1 0 1 0 1
55 1 1 0 1 1 0
56 1 1 0 1 1 1
57 1 1 1 0 0 0
58 1 1 1 0 0 1
59 1 1 1 0 1 0
60 1 1 1 0 1 1
61 1 1 1 1 0 0
62 1 1 1 1 0 1
63 1 1 1 1 1 0
64 1 1 1 1 1 1
;
run;
proc sort;by rp;run;

%macro createds (index=1);
data g.cvarpats;set g.cvarpats;

e5m4=0;
e5sub=0;

/* HEteroG R-1 (Triplet) AGREEMENT *//* HOmoG RE: CATEGORY */
      R12345=0; R12346=0; R12356=0; R12456=0;R13456=0; R23456=0;

      if R1=R2 and R1=R3 and  R1=R4 and R1=R5 and R6~=R1 then R12345=1;
      if R12345=1 then e5sub=1;

      if R1=R2 and R1=R3 and  R1=R4 and R1=R6 and R5~=R1 then R12346=1;
      if R12346=1 then e5sub=1;

      if R1=R2 and R1=R3 and  R1=R6 and R1=R5 and R4~=R1 then R12356=1;
      if R12356=1 then e5at=1;

      if R1=R2 and R1=R6 and  R1=R4 and R1=R5 and R3~=R1 then R12456=1;
      if R12456=1 then e5sub=1;

      if R1=R6 and R1=R3 and  R1=R4 and R1=R5 and R2~=R1 then R13456=1;
      if R13456=1 then e5sub=1;

      if R6=R2 and R6=R3 and  R6=R4 and R6=R5 and R1~=R2 then R23456=1;
      if R23456=1 then e5sub=1;
```

```
e6=0;e6c0=0; e6c1=0;
if R1=R2 and R1=R3 and R1=R4 and R1=R5 and R1=R6 then e6=1;
if R1=R2 and R1=R3 and R1=R4 and R1=R5 and R1=R6 and R1=0 then e6c0=1;
if R1=R2 and R1=R3 and R1=R4 and R1=R5 and R1=R6 and R1=1 then e6c1=1;
/* Homogeneous with respect to category between R2, R3, and R4 only */
e5m1=0; e5m2=0;e5m3=0;e5m4=0;e5m5=0;e5m6=0;
e5c0=0; e5c1=0;
if R12345=1 and R1=0 then e5c0=1; if R12345=1 and R1=1 then e5c1=1;
if R12346=1 and R1=0 then e5c0=1; if R12346=1 and R1=1 then e5c1=1;
if R12356=1 and R1=0 then e5c0=1; if R12356=1 and R1=1 then e5c1=1;
if R12456=1 and R1=0 then e5c0=1; if R12456=1 and R1=1 then e5c1=1;
if R13456=1 and R1=0 then e5c0=1; if R13456=1 and R1=1 then e5c1=1;
if R23456=1 and R2=0 then e5c0=1; if R23456=1 and R2=1 then e5c1=1;

if R12345=1 then e5m6=1;
if R12346=1 then e5m5=1;
if R12356=1 then e5m4=1;
if R12456=1 then e5m3=1;
if R13456=1 then e5m2=1;
if R23456=1 then e5m1=1;
e5=0;
if e5m6=1 or e5m5=1 or e5m4=1 or e5m3=1 or e5m2=1 or e5m1=1 then e5=1;
/*need rating of zero as -1
because of negative one, one parameterization*/
a=r1;b=r2;c=r3;d=r4;e=r5;f=r6;
if r1=0 then r1=-1;if r2=0 then r2=-1;if r3=0 then r3=-1;
if r4=0 then r4=-1;if r5=0 then r5=-1;if r6=0 then r6=-1;
run;


data Bvector;
 set get.atyp_1k;
 /*Data set containg Beta vector of
 1,000 simulated GHeP Moderate Model under
 the assumption of Marginal HOMOGENEITY*/ run;

data Bvector;
 do TAKEIT=1 to 1000 BY 1;
 /* 1 to the Number of SIMULATIONS done, here 1,000*/
 set Bvector POINT=TAKEIT;
 simN=takeit;
 output;
 end;
 stop;
run;



data ds&index;
 set Bvector;
  do i=1 to 64;
  rp=i;
  if simN=&index then output;
  end;
  run;
```

```
proc sort data=ds&index;by rp;
proc sort data=g.cvarpats;by rp;

  data ds&index (drop=cntN&index COL1-COL4);
  merge ds&index g.cvarpats; by rp;

  /*d6=col1       e5sub=col2          e5m4=col3        mu=col4
3.575551  0.9628107  1.966113     -0.8675006*/


  b1=COL1; b2=COL2; b3=COL3; mu4=COL4;

/* For each dataset (ds#) the values of b1-b3 and mu4 are set to
the values generated from the corresponding simulation #. For each
rating (covariate) pattern, the value of the variable logm is computed
from the corresponding sum of the appropriate parameter estimates.

For the first simulation (&index=1), The variable cntN1 is calculated
by exponentiating the sum of the parameter estimates corresponding that
each rating pattern.  The count data for a given rating pattern is
computed by randomly sampling from a Poisson distribution with a mean
equal to the value of the variable cntN1.

The rating pattern is constructed by concatenating the values of r1
through
R6 and removing the any internal spaces (compress function).

This algorithm is repeated for the other 999 simulations. */

logm=mu4+e6*b1+e5sub*b2+e5m4*b3;
cntN&index=exp(logm);
smcnt&index = ranpoi(0,cntN&index);
cnt&index=round(smcnt&index);
pattern=trim(a)||trim(b)||trim(c)||trim(d)||trim(e)||trim(f);
pattern=compress(pattern);
run;
proc sort;by pattern;run;

data stat&index (keep=rp cnt&index);
 set ds&index;

proc transpose data=stat&index out=ssf&index prefix=rpcnt;
 id rp;
 var cnt&index;run;

data g.suffstat;set g.suffstat ssf&index;run;
%mend createds;


/*One - One Hundred*/
%createds(index=1); %createds(index=2);%createds(index=3);
%createds(index=4);%createds(index=5);
%createds(index=6);%createds(index=7);
%createds(index=8); %createds(index=9);
%createds(index=10);
:
```

```
:
:%createds(index=995);%createds(index=996);%createds(index=997);
%createds(index=998); %createds(index=999); %createds(index=1000);
run;


data g.suffstat (drop=what);
 set g.suffstat;
 id=_N_;
 if id=1 then delete;
 run;

data g.suffstat;
 set g.suffstat;
 id=id-1;
 total=sum(of rpcnt1-rpcnt64);

/*Variables to determine if sufficient statistic for
heterogeneous partial agreement parameter is zero*/
ssm1=1;ssm2=1;ssm3=1;ssm4=1;ssm5=1;ssm6=1;
if (rpcnt2=0) and (rpcnt63=0) then ssm6=0;
if (rpcnt3=0) and (rpcnt62=0) then ssm5=0;
if (rpcnt5=0) and (rpcnt60=0) then ssm4=0;
if (rpcnt9=0) and (rpcnt56=0) then ssm3=0;
if (rpcnt17=0) and (rpcnt48=0) then ssm2=0;
if (rpcnt32=0) and (rpcnt33=0) then ssm1=0;

sufst=ssm1+ssm2+ssm3+ssm4+ssm5+ssm6;
*if what='suffstat' then sufst=.;
/*No raters whose partial agreement cnt = zero*/
if (ssm1=1 and ssm2=1 and ssm3=1 and ssm4=1 and ssm5=1 and ssm6=1)
then model=1;
/*one raters whose partial agreement cnt = zero*/
if (ssm1=1 and ssm2=1 and ssm3=1 and ssm4=1 and ssm5=1 and ssm6=0)
then model=2;
if (ssm1=1 and ssm2=1 and ssm3=1 and ssm4=1 and ssm5=0 and ssm6=1)
then model=3;
if (ssm1=1 and ssm2=1 and ssm3=1 and ssm4=0 and ssm5=1 and ssm6=1)
then model=4;
if (ssm1=1 and ssm2=1 and ssm3=0 and ssm4=1 and ssm5=1 and ssm6=1)
then model=5;
if (ssm1=1 and ssm2=0 and ssm3=1 and ssm4=1 and ssm5=1 and ssm6=1)
then model=6;
if (ssm1=0 and ssm2=1 and ssm3=1 and ssm4=1 and ssm5=1 and ssm6=1)
then model=7;

/*Two raters whose partial agreement cnt = zero*/
if (ssm1=0 and ssm2=0 and ssm3=1 and ssm4=1 and ssm5=1 and ssm6=1)
then model=8;
if (ssm1=0 and ssm2=1 and ssm3=0 and ssm4=1 and ssm5=1 and ssm6=1)
then model=9;
if (ssm1=0 and ssm2=1 and ssm3=1 and ssm4=0 and ssm5=1 and ssm6=1)
then model=10;
if (ssm1=0 and ssm2=1 and ssm3=1 and ssm4=1 and ssm5=0 and ssm6=1)
then model=11;
if (ssm1=0 and ssm2=1 and ssm3=1 and ssm4=1 and ssm5=1 and ssm6=0)
then model=12;
```

```
if (ssm1=1 and ssm2=0 and ssm3=0 and ssm4=1 and ssm5=1 and ssm6=1)
then model=13;
if (ssm1=1 and ssm2=0 and ssm3=1 and ssm4=0 and ssm5=1 and ssm6=1)
then model=14;
if (ssm1=1 and ssm2=0 and ssm3=1 and ssm4=1 and ssm5=0 and ssm6=1)
then model=15;
if (ssm1=1 and ssm2=0 and ssm3=1 and ssm4=1 and ssm5=1 and ssm6=0)
then model=16;
if (ssm1=1 and ssm2=1 and ssm3=0 and ssm4=0 and ssm5=1 and ssm6=1)
then model=17;
if (ssm1=1 and ssm2=1 and ssm3=0 and ssm4=1 and ssm5=0 and ssm6=1)
then model=18;
if (ssm1=1 and ssm2=1 and ssm3=0 and ssm4=1 and ssm5=1 and ssm6=0)
then model=19;
if (ssm1=1 and ssm2=1 and ssm3=1 and ssm4=0 and ssm5=0 and ssm6=1)
then model=20;
if (ssm1=1 and ssm2=1 and ssm3=1 and ssm4=0 and ssm5=1 and ssm6=0)
then model=21;
if (ssm1=1 and ssm2=1 and ssm3=1 and ssm4=1 and ssm5=0 and ssm6=0)
then model=22;


/*Three raters whose partial agreement cnt = zero*/
if (ssm1=0 and ssm2=0 and ssm3=0 and ssm4=1 and ssm5=1 and ssm6=1)
then model=23;
if (ssm1=0 and ssm2=0 and ssm3=1 and ssm4=0 and ssm5=1 and ssm6=1)
then model=24;
if (ssm1=0 and ssm2=0 and ssm3=1 and ssm4=1 and ssm5=0 and ssm6=1)
then model=25;
if (ssm1=0 and ssm2=0 and ssm3=1 and ssm4=1 and ssm5=1 and ssm6=0)
then model=26;
if (ssm1=1 and ssm2=0 and ssm3=0 and ssm4=0 and ssm5=1 and ssm6=1)
then model=27;
if (ssm1=1 and ssm2=0 and ssm3=0 and ssm4=1 and ssm5=0 and ssm6=1)
then model=28;
if (ssm1=1 and ssm2=0 and ssm3=0 and ssm4=1 and ssm5=1 and ssm6=0)
then model=29;
if (ssm1=1 and ssm2=1 and ssm3=0 and ssm4=0 and ssm5=0 and ssm6=1)
then model=30;
if (ssm1=1 and ssm2=1 and ssm3=0 and ssm4=0 and ssm5=1 and ssm6=0)
then model=31;
if (ssm1=0 and ssm2=1 and ssm3=0 and ssm4=0 and ssm5=1 and ssm6=1)
then model=32;
if (ssm1=0 and ssm2=1 and ssm3=0 and ssm4=1 and ssm5=0 and ssm6=1)
then model=33;
if (ssm1=0 and ssm2=1 and ssm3=0 and ssm4=1 and ssm5=1 and ssm6=0)
then model=34;
if (ssm1=0 and ssm2=1 and ssm3=1 and ssm4=0 and ssm5=0 and ssm6=1)
then model=35;
if (ssm1=0 and ssm2=1 and ssm3=1 and ssm4=1 and ssm5=0 and ssm6=0)
then model=36;
if (ssm1=0 and ssm2=1 and ssm3=1 and ssm4=0 and ssm5=1 and ssm6=0)
then model=37;
if (ssm1=1 and ssm2=0 and ssm3=1 and ssm4=0 and ssm5=0 and ssm6=1)
then model=38;
if (ssm1=1 and ssm2=0 and ssm3=1 and ssm4=0 and ssm5=1 and ssm6=0)
then model=39;
```

```
if (ssm1=1 and ssm2=1 and ssm3=1 and ssm4=0 and ssm5=0 and ssm6=0)
then model=40;
if (ssm1=1 and ssm2=1 and ssm3=0 and ssm4=1 and ssm5=0 and ssm6=0)
then model=41;
if (ssm1=1 and ssm2=0 and ssm3=1 and ssm4=1 and ssm5=0 and ssm6=0)
then model=42;

/*four raters whose partial agreement cnt = zero*/
if (ssm1=1 and ssm2=1 and ssm3=0 and ssm4=0 and ssm5=0 and ssm6=0)
then model=43;
if (ssm1=1 and ssm2=0 and ssm3=1 and ssm4=0 and ssm5=0 and ssm6=0)
then model=44;
if (ssm1=1 and ssm2=0 and ssm3=0 and ssm4=1 and ssm5=0 and ssm6=0)
then model=45;
if (ssm1=1 and ssm2=0 and ssm3=0 and ssm4=0 and ssm5=1 and ssm6=0)
then model=46;
if (ssm1=1 and ssm2=0 and ssm3=0 and ssm4=0 and ssm5=0 and ssm6=1)
then model=47;
if (ssm1=0 and ssm2=1 and ssm3=1 and ssm4=0 and ssm5=0 and ssm6=0)
then model=48;
if (ssm1=0 and ssm2=1 and ssm3=0 and ssm4=1 and ssm5=0 and ssm6=0)
then model=49;
if (ssm1=0 and ssm2=1 and ssm3=0 and ssm4=0 and ssm5=1 and ssm6=0)
then model=50;
if (ssm1=0 and ssm2=1 and ssm3=0 and ssm4=0 and ssm5=0 and ssm6=1)
then model=51;
if (ssm1=0 and ssm2=0 and ssm3=1 and ssm4=1 and ssm5=0 and ssm6=0)
then model=52;
if (ssm1=0 and ssm2=0 and ssm3=1 and ssm4=0 and ssm5=1 and ssm6=0)
then model=53;
if (ssm1=0 and ssm2=0 and ssm3=1 and ssm4=0 and ssm5=0 and ssm6=1)
then model=54;
if (ssm1=0 and ssm2=0 and ssm3=0 and ssm4=1 and ssm5=1 and ssm6=0)
then model=55;
if (ssm1=0 and ssm2=0 and ssm3=0 and ssm4=1 and ssm5=0 and ssm6=1)
then model=56;
if (ssm1=0 and ssm2=0 and ssm3=0 and ssm4=0 and ssm5=1 and ssm6=1)
then model=57;



/*five raters whose partial agreement cnt = zero*/
if (ssm1=1 and ssm2=0 and ssm3=0 and ssm4=0 and ssm5=0 and ssm6=0)
then model=58;
if (ssm1=0 and ssm2=1 and ssm3=0 and ssm4=0 and ssm5=0 and ssm6=0)
then model=59;
if (ssm1=0 and ssm2=0 and ssm3=1 and ssm4=0 and ssm5=0 and ssm6=0)
then model=60;
if (ssm1=0 and ssm2=0 and ssm3=0 and ssm4=1 and ssm5=0 and ssm6=0)
then model=61;
if (ssm1=0 and ssm2=0 and ssm3=0 and ssm4=0 and ssm5=1 and ssm6=0)
then model=62;
if (ssm1=0 and ssm2=0 and ssm3=0 and ssm4=0 and ssm5=0 and ssm6=1)
then model=63;
if (ssm1=0 and ssm2=0 and ssm3=0 and ssm4=0 and ssm5=0 and ssm6=0)
then model=64;
run;
```

```
proc print data=g.suffstat;title1 'Step F';run;
proc freq data=g.suffstat;table model ;
title1 'RPM  GHeP MHomoG: Model # re: # of Suff Stats=0';
 run;

 proc freq data=g.suffstat;table total;
title1 'RME GHeP MHomoG: Sample Size';
 run;


data G1_w      G2_w      G3_w        G4_w
     G5_w      G6_w      G7_w        G8_w
     G9_w      G11_w     G12_w
     G13_w     G15_w     G16_w       G17_w
     G18_w     G19_w
     G22_w     G23_w     G25_w       G26_w
     G28_w     G29_w     G33_w
     G34_w     G36_w     G39_w       G41_w       G42_w
     G45_w     G49_w     G52_w
     G55_w     G56_w     G61_w      ;

 set g.suffstat ;
 if model=1 then output G1_w;
 else if model=2 then output G2_w;
 else if model=3 then output G3_w;
 else if model=4 then output G4_w;
 else if model=5 then output G5_w;
 else if model=6 then output G6_w;
 else if model=7 then output G7_w;
 else if model=8 then output G8_w;
 else if model=9 then output G9_w;
 *else if model=10 then output G10_w;
 else if model=11 then output G11_w;
 else if model=12 then output G12_w;
 else if model=13 then output G13_w;
 *else if model=14 then output G14_w;
 else if model=15 then output G15_w;
 else if model=16 then output G16_w;
 else if model=17 then output G17_w;
 else if model=18 then output G18_w;
 else if model=19 then output G19_w;
 *else if model=20 then output G20_w;
 *else if model=21 then output G21_w;
 else if model=22 then output G22_w;
 else if model=23 then output G23_w;
* else if model=24 then output G24_w;
else if model=25 then output G25_w;
else if model=26 then output G26_w;
*else if model=27 then output G27_w;
else if model=28 then output G28_w;
else if model=29 then output G29_w;
*else if model=30 then output G30_w;
*else if model=31 then output G31_w;
*else if model=32 then output G32_w;
else if model=33 then output G33_w;
else if model=34 then output G34_w;
*else if model=35 then output G35_w;
```

```
else if model=36 then output G36_w;
*else if model=37 then output G37_w;
*else if model=38 then output G38_w;
else if model=39 then output G39_w;
*else if model=40 then output G40_w;
else if model=41 then output G41_w;
else if model=42 then output G42_w;
*else if model=43 then output G43_w;
*else if model=44 then output G44_w;
else if model=45 then output G45_w;
*else if model=46 then output G46_w;
*else if model=47 then output G47_w;
*else if model=48 then output G48_w;
else if model=49 then output G49_w;
*else if model=50 then output G50_w;
*else if model=51 then output G51_w;
else if model=52 then output G52_w;
*else if model=53 then output G53_w;
*else if model=54 then output G54_w;
else if model=55 then output G55_w;
else if model=56 then output G56_w;
*else if model=57 then output G57_w;
*else if model=58 then output G58_w;
*else if model=59 then output G59_w;
*else if model=60 then output G60_w;
else if model=61 then output G61_w;
*else if model=62 then output G62_w;
*else if model=63 then output G63_w;
*else if model=64 then output G64_w;

/*Model 1*/
data G1_w ( drop=ssm1-ssm6 sufst model id);  set G1_w; run;
proc transpose data=G1_w out=G1_long; run;
data g.G1 (drop=_NAME_); set G1_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G1;by rp;run;
/*Model 2*/
data G2_w ( drop=ssm1-ssm6 sufst model id);  set G2_w; run;
proc transpose data=G2_w out=G2_long; run;
data g.G2 (drop=_NAME_); set G2_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G2;by rp;run;

/*Model 3*/
data G3_w ( drop=ssm1-ssm6 sufst model id);  set G3_w; run;
proc transpose data=G3_w out=G3_long; run;
data g.G3 (drop=_NAME_); set G3_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G3;by rp;run;
/*Model 4*/
data G4_w ( drop=ssm1-ssm6 sufst model id);  set G4_w; run;
proc transpose data=G4_w out=G4_long; run;
data g.G4 (drop=_NAME_); set G4_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G4;by rp;run;

/*Model 5*/
data G5_w ( drop=ssm1-ssm6 sufst model id);  set G5_w; run;
proc transpose data=G5_w out=G5_long; run;
data g.G5 (drop=_NAME_); set G5_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G5;by rp;run;
```

```
/*Model 6*/
data G6_w ( drop=ssm1-ssm6 sufst model id);  set G6_w; run;
proc transpose data=G6_w out=G6_long; run;
data g.G6 (drop=_NAME_); set G6_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G6;by rp;run;

/*Model 7*/
data G7_w ( drop=ssm1-ssm6 sufst model id);  set G7_w; run;
proc transpose data=G7_w out=G7_long; run;
data g.G7 (drop=_NAME_); set G7_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G7;by rp;run;
/*Model 8*/
data G8_w ( drop=ssm1-ssm6 sufst model id);  set G8_w; run;
proc transpose data=G8_w out=G8_long; run;
data g.G8 (drop=_NAME_); set G8_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G8;by rp;run;

/*Model 9*/
data G9_w ( drop=ssm1-ssm6 sufst model id);  set G9_w; run;
proc transpose data=G9_w out=G9_long; run;
data g.G9 (drop=_NAME_); set G9_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G9;by rp;run;


/*Model 11*/
data G11_w ( drop=ssm1-ssm6 sufst model id);  set G11_w; run;
proc transpose data=G11_w out=G11_long; run;
data g.G11 (drop=_NAME_); set G11_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G11;by rp;run;

/*Model 12*/
data G12_w ( drop=ssm1-ssm6 sufst model id);  set G12_w; run;
proc transpose data=G12_w out=G12_long; run;
data g.G12 (drop=_NAME_); set G12_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G12;by rp;run;
/*Model 13*/
data G13_w ( drop=ssm1-ssm6 sufst model id);  set G13_w; run;
proc transpose data=G13_w out=G13_long; run;
data g.G13 (drop=_NAME_); set G13_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G13;by rp;run;

/*Model 15*/
data G15_w ( drop=ssm1-ssm6 sufst model id);  set G15_w; run;
proc transpose data=G15_w out=G15_long; run;
data g.G15 (drop=_NAME_); set G15_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G15;by rp;run;
/*Model 16*/
data G16_w ( drop=ssm1-ssm6 sufst model id);  set G16_w; run;
proc transpose data=G16_w out=G16_long; run;
data g.G16 (drop=_NAME_); set G16_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G16;by rp;run;
/*Model 17*/
data G17_w ( drop=ssm1-ssm6 sufst model id);  set G17_w; run;
proc transpose data=G17_w out=G17_long; run;
data g.G17 (drop=_NAME_); set G17_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G17;by rp;run;
```

```
/*Model 18*/
data G18_w ( drop=ssm1-ssm6 sufst model id);  set G18_w; run;
proc transpose data=G18_w out=G18_long; run;
data g.G18 (drop=_NAME_); set G18_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G18;by rp;run;

/*Model 19*/
data G19_w ( drop=ssm1-ssm6 sufst model id);  set G19_w; run;
proc transpose data=G19_w out=G19_long; run;
data g.G19 (drop=_NAME_); set G19_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G19;by rp;run;


/*Model 22*/
data G22_w ( drop=ssm1-ssm6 sufst model id);  set G22_w; run;
proc transpose data=G22_w out=G22_long; run;
data g.G22 (drop=_NAME_); set G22_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G22;by rp;run;

/*Model 23*/
data G23_w ( drop=ssm1-ssm6 sufst model id);  set G23_w; run;
proc transpose data=G23_w out=G23_long; run;
data g.G23 (drop=_NAME_); set G23_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G23;by rp;run;

/*Model 25*/
data G25_w ( drop=ssm1-ssm6 sufst model id);  set G25_w; run;
proc transpose data=G25_w out=G25_long; run;
data g.G25 (drop=_NAME_); set G25_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G25;by rp;run;


/*Model 26*/
data G26_w ( drop=ssm1-ssm6 sufst model id);  set G26_w; run;
proc transpose data=G26_w out=G26_long; run;
data g.G26 (drop=_NAME_); set G26_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G26;by rp;run;

/*Model 28*/
data G28_w ( drop=ssm1-ssm6 sufst model id);  set G28_w; run;
proc transpose data=G28_w out=G28_long; run;
data g.G28 (drop=_NAME_); set G28_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G28;by rp;run;

/*Model 29*/
data G29_w ( drop=ssm1-ssm6 sufst model id);  set G29_w; run;
proc transpose data=G29_w out=G29_long; run;
data g.G29 (drop=_NAME_); set G29_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G29;by rp;run;


/*Model 33*/
data G33_w ( drop=ssm1-ssm6 sufst model id);  set G33_w; run;
proc transpose data=G33_w out=G33_long; run;
data g.G33 (drop=_NAME_); set G33_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G33;by rp;run;
```

```
/*Model 34*/
data G34_w ( drop=ssm1-ssm6 sufst model id);  set G34_w; run;
proc transpose data=G34_w out=G34_long; run;
data g.G34 (drop=_NAME_); set G34_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G34;by rp;run;

/*Model 36*/
data G36_w ( drop=ssm1-ssm6 sufst model id);  set G36_w; run;
proc transpose data=G36_w out=G36_long; run;
data g.G36 (drop=_NAME_); set G36_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G36;by rp;run;

/*Model 39*/
data G39_w ( drop=ssm1-ssm6 sufst model id);  set G39_w; run;
proc transpose data=G39_w out=G39_long; run;
data g.G39 (drop=_NAME_); set G39_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G39;by rp;run;


/*Model 41*/
data G41_w ( drop=ssm1-ssm6 sufst model id);  set G41_w; run;
proc transpose data=G41_w out=G41_long; run;
data g.G41 (drop=_NAME_); set G41_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G41;by rp;run;


/*Model 42*/
data G42_w ( drop=ssm1-ssm6 sufst model id);  set G42_w; run;
proc transpose data=G42_w out=G42_long; run;
data g.G42 (drop=_NAME_); set G42_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G42;by rp;run;


/*Model 45*/
data G45_w ( drop=ssm1-ssm6 sufst model id);  set G45_w; run;
proc transpose data=G45_w out=G45_long; run;
data g.G45 (drop=_NAME_); set G45_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G45;by rp;run;



/*Model 49*/
data G49_w ( drop=ssm1-ssm6 sufst model id);  set G49_w; run;
proc transpose data=G49_w out=G49_long; run;
data g.G49 (drop=_NAME_); set G49_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G49;by rp;run;


/*Model 52*/
data G52_w ( drop=ssm1-ssm6 sufst model id);  set G52_w; run;
proc transpose data=G52_w out=G52_long; run;
data g.G52 (drop=_NAME_); set G52_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G52;by rp;run;

/*Model 55*/
data G55_w ( drop=ssm1-ssm6 sufst model id);  set G55_w; run;
proc transpose data=G55_w out=G55_long; run;
```

```
data g.G55 (drop=_NAME_); set G55_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G55;by rp;run;

/*Model 56*/
data G56_w ( drop=ssm1-ssm6 sufst model id);  set G56_w; run;
proc transpose data=G56_w out=G56_long; run;
data g.G56 (drop=_NAME_); set G56_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G56;by rp;run;

/*Model 61*/
data G61_w ( drop=ssm1-ssm6 sufst model id);  set G61_w; run;
proc transpose data=G61_w out=G61_long; run;
data g.G61 (drop=_NAME_); set G61_long; rp=substr(_NAME_,6)+0;run;
proc sort data=g.G61;by rp;run;


/* datasets with neg one/one parameterization */

data g.G1;merge g.cvarpats g.G1;by rp;run;
data g.G2;merge g.cvarpats g.G2;by rp;run;
data g.G3;merge g.cvarpats g.G3;by rp;run;
data g.G4;merge g.cvarpats g.G4;by rp;run;
data g.G5;merge g.cvarpats g.G5;by rp;run;
data g.G6;merge g.cvarpats g.G6;by rp;run;
data g.G7;merge g.cvarpats g.G7;by rp;run;
data g.G8;merge g.cvarpats g.G8;by rp;run;
data g.G9;merge g.cvarpats g.G9;by rp;run;

data g.G11;merge g.cvarpats g.G11;by rp;run;
data g.G12;merge g.cvarpats g.G12;by rp;run;
data g.G13;merge g.cvarpats g.G13;by rp;run;

data g.G15;merge g.cvarpats g.G15;by rp;run;
data g.G16;merge g.cvarpats g.G16;by rp;run;
data g.G17;merge g.cvarpats g.G17;by rp;run;
data g.G18;merge g.cvarpats g.G18;by rp;run;
data g.G19;merge g.cvarpats g.G19;by rp;run;

data g.G21;merge g.cvarpats g.G21;by rp;run;
data g.G22;merge g.cvarpats g.G22;by rp;run;
data g.G23;merge g.cvarpats g.G23;by rp;run;

data g.G25;merge g.cvarpats g.G25;by rp;run;
data g.G26;merge g.cvarpats g.G26;by rp;run;

data g.G28;merge g.cvarpats g.G28;by rp;run;
data g.G29;merge g.cvarpats g.G29;by rp;run;



data g.G33;merge g.cvarpats g.G33;by rp;run;
data g.G34;merge g.cvarpats g.G34;by rp;run;

data g.G36;merge g.cvarpats g.G36;by rp;run;
data g.G39;merge g.cvarpats g.G39;by rp;run;

data g.G41;merge g.cvarpats g.G41;by rp;run;
data g.G42;merge g.cvarpats g.G42;by rp;run;

data g.G45;merge g.cvarpats g.G45;by rp;run;


data g.G49;merge g.cvarpats g.G49;by rp;run;

data g.G52;merge g.cvarpats g.G52;by rp;run;

data g.G55;merge g.cvarpats g.G55;by rp;run;
data g.G56;merge g.cvarpats g.G56;by rp;run;

data g.G61;merge g.cvarpats g.G61;by rp;run;
```

Appendix F
STATA Program Used to Perform Pair-wise Comparisons of the GHeP paramaters

```
/**********************************************************/
/*pw_G1.do                                        */
/*Dissertation                                            */
/*Conducts pairwise comparisons of the heterogeneous      */
/* partial agreement parameters for dataset               */
/*g#.dta                                  */
/*Simulations - pw done using lincom command              */
/*Captures the estimate, se and df for the G Model      */
/*for each simulation.                                    */
/* K.B. Kastango                                          */
/**********************************************************/

capture program drop pw_G1
program define pw_G1, rclass
       /*  Version 8.0*/

capture log close
log using
"C:\aaPhDSimulations\HomoG\G\adofiles\do_pwG1_pval.log",replace
use "C:\aaPhDSimulations\HomoG\G\STATAds\G1.dta", clear
display "Opened Data Set G1.dta - X sims, e5m=0"

drop if rp == .

quietly {
 generate rpnum=_n
}
quietly {
foreach x of varlist cnt* {

/*Fit GHeP Model, create var estB & seEst*/
                glm `x' e6 e5m1 e5m2 e5m3 e5m4 e5m5 e5m6, f(p)

                lincom e5m1-e5m2
                gen est12z`x'=r(estimate)
                gen se12z`x'=r(se)
                gen z12a`x'=est12z`x'/se12z`x'

                lincom e5m1-e5m3
                gen est13z`x'=r(estimate)
                gen se13z`x'=r(se)
                gen z13a`x'=est13z`x'/se13z`x'

                lincom e5m1-e5m4
                gen est14z`x'=r(estimate)
                gen se14z`x'=r(se)
                gen z14a`x'=est14z`x'/se14z`x'


                lincom e5m1-e5m5
                gen est15z`x'=r(estimate)
                gen se15z`x'=r(se)
                gen z15a`x'=est15z`x'/se15z`x'
                lincom e5m1-e5m6
                gen est16z`x'=r(estimate)
                gen se16z`x'=r(se)
                gen z16a`x'=est16z`x'/se16z`x'

                lincom e5m2-e5m3
                gen est23z`x'=r(estimate)
                gen se23z`x'=r(se)
                gen z23a`x'=est23z`x'/se23z`x'

                lincom e5m2-e5m4
                gen est24z`x'=r(estimate)
                gen se24z`x'=r(se)
                gen z24a`x'=est24z`x'/se24z`x'

                lincom e5m2-e5m5
                gen est25z`x'=r(estimate)
                gen se25z`x'=r(se)
                gen z25a`x'=est25z`x'/se25z`x'

                lincom e5m2-e5m6
                gen est26z`x'=r(estimate)
                gen se26z`x'=r(se)
                gen z26a`x'=est26z`x'/se26z`x'

                lincom e5m3-e5m4
                gen est34z`x'=r(estimate)
                gen se34z`x'=r(se)
                gen z34a`x'=est34z`x'/se34z`x'


                lincom e5m3-e5m5
                gen est35z`x'=r(estimate)
                gen se35z`x'=r(se)
                gen z35a`x'=est35z`x'/se35z`x'

                lincom e5m3-e5m6
                gen est36z`x'=r(estimate)
                gen se36z`x'=r(se)
                gen z36a`x'=est36z`x'/se36z`x'

               lincom e5m4-e5m5
                gen est45z`x'=r(estimate)
                gen se45z`x'=r(se)
                gen z45a`x'=est45z`x'/se45z`x'

               lincom e5m4-e5m6
                gen est46z`x'=r(estimate)
                gen se46z`x'=r(se)
                gen z46a`x'=est46z`x'/se46z`x'

                lincom e5m5-e5m6
                gen est56z`x'=r(estimate)
                gen se56z`x'=r(se)
                gen z56a`x'=est56z`x'/se56z`x'

                                      }
```

```
                                      }

quietly {
   drop if rp>1
   drop r1 r2 r3 r4 r5 r6
   drop e6 e6c0 e6c1 e5m1 e5m2
   drop e5m3 e5m4 e5m5 e5m6 e5c0 e5c1 e5
   drop cnt*
}
save
"C:\aaPhDSimulations\HomoG\G\STATAds\coverage\ado_homg_G1_pval.dta"
save "C:\aaPhDSimulations\HomoG\G\STATAds\pwfile\ado_pw_G1_pval.dta"

use "C:\aaPhDSimulations\HomoG\G\STATAds\pwfile\ado_pw_G1_pval.dta",
clear
display "Opened Data Set"
drop est*  se*

reshape long z12acnt z13acnt z14acnt z15acnt z16acnt z23acnt z24acnt
z25acnt z26acnt z34acnt z35acnt z36acnt z45acnt z46acnt z56acnt, i(rp)
j(sim)

replace z12acnt = -1*abs(z12acnt)
replace z13acnt = -1*abs(z13acnt)
replace z14acnt = -1*abs(z14acnt)
replace z15acnt = -1*abs(z15acnt)
replace z16acnt = -1*abs(z16acnt)
replace z23acnt = -1*abs(z23acnt)
replace z24acnt = -1*abs(z24acnt)
replace z25acnt = -1*abs(z25acnt)
replace z26acnt = -1*abs(z26acnt)
replace z34acnt = -1*abs(z34acnt)
replace z35acnt = -1*abs(z35acnt)
replace z36acnt = -1*abs(z36acnt)
replace z45acnt = -1*abs(z45acnt)
replace z46acnt = -1*abs(z46acnt)
replace z56acnt = -1*abs(z56acnt)

gen p12=norm(z12acnt)
gen p13=norm(z13acnt)
gen p14=norm(z14acnt)
gen p15=norm(z15acnt)
gen p16=norm(z16acnt)
gen p23=norm(z23acnt)
gen p24=norm(z24acnt)
gen p25=norm(z25acnt)
gen p26=norm(z26acnt)
gen p34=norm(z34acnt)
gen p35=norm(z35acnt)
gen p36=norm(z36acnt)
gen p45=norm(z45acnt)
gen p46=norm(z46acnt)
gen p56=norm(z56acnt)


save
"C:\aaPhDSimulations\HomoG\G\STATAds\pwfile\ado_reshape_pw_G1_pval.dta"
```

```
use
"C:\aaPhDSimulations\HomoG\G\STATAds\pwfile\ado_reshape_pw_G1_pval.dta"
, clear

drop  rp rpnum z*


reshape long p, i(sim) j(hyp)

gen p2=p*2
gen ncomp=15
gen Atyp=0
replace Atyp=1 if p2 <= 0.0034
list sim hyp p2 if Atyp==1
table sim Atyp
gen Btyp=0
replace Btyp=1 if p2 <=0.05
list sim hyp p2 if Btyp==1
save "C:\aaPhDSimulations\HomoG\G\STATAds\pwfile\finalpval_G1.dta"
capture log close

end
pw_G1
```

Appendix G
SAS Code to Perform Multiple Comparison Procedures

```
/*This program is centered around the SAS procedure PROC MULTTEST*/

/* Dataset "pvals" contains the p-value of the 15, 10, 6, 3, or 1
possible multiple comparisons of the heterogeneous K-1 partial
agreement parameters and a character variable denoting which pair-wise
comparison the p-value is from. P-values of each comparison was
determined using STATA (see Appendix F)*/

/* For example, H12 0.610 indicates that the p-value of the pair-wise
comparison between the heterogeneous partial agreement parameter of
raters 1 and 2.

Ho: e5m1 – e5m2 =0 ;
*/


/* This specific program is for simulations that have sufficient
statistics equal to zero as described by model #56 */

libname get "C:\aaPhDSimulations\HomoG\GHePSevere\STATAds\pwfile";
libname posthoc
"C:\aaPhDSimulations\HomoG\RogSimData\multtest_results";


 proc contents data=get.finalpval_g56;run; proc sort
data=get.finalpval_g56;by sim;run;
/*proc print;title1 'Model 56 ROG_SIM_DATA';run;*/
data mod56sims;set get.finalpval_g56;by sim;if first.sim;run;
/*
proc print;run;
 data mod56sims (keep=adjc resultc sim);
 set mod56sims;
contrast='H'||trim(hyp);
adjc='%adj56(sim='||trim(sim)||');';
adjc=compress(adjc);
resultc='posthoc.results'||trim(sim)||';';
resultc=compress(resultc);
data adj (keep=adjc) result(keep=resultc);
 set mod56sims;run;
 proc print data=adj noobs;title1 'Model 56 ROG_SIM_DATA';run;
 proc print data=result noobs;run;
*/
%macro adj56(sim=0);
data Sev&sim (rename=(p2=raw_p));
 set get.finalpval_g56;
 contrast='H'||trim(hyp);
 contrast= compress(contrast);
 where sim=&sim;
run;
proc multtest pdata=Sev&sim bon sid holm stepsid fdr
out=posthoc.results&sim;
title1 'Bonferroni, Sidak, Stepdown Bon, Stepdown Sidak, False
Discovery Rate';
title2 "Simulation Scenario &sim ROG_SIM_DATA";run;
proc print data=posthoc.results&sim;run;
%mend adj56;
              %adj56(sim=791); %adj56(sim=977);
```

```
data model56 (keep=sim hyp bon_p);
 set  posthoc.results791 posthoc.results977 ;
  run;proc sort;by sim hyp;
data model56;  set model56;    n=_n_;     run;
proc transpose out=first(drop=_name_); by n sim; var hyp bon_p ; run;
proc transpose data=first out=posthoc.m56Bon(drop=_name_) prefix=pbon;
by sim;   var col1;   proc print;
title1 'Bonferroni - Model 56 ROG_SIM_DATA';
title2 ' ' ;run;

data posthoc.m56bon (rename=(pbon2=pbonN));
 set posthoc.m56bon;run;
 data posthoc.m56bon (drop=pbon1 );
 set posthoc.m56bon;
model=56;run;
 proc print;
title1 ' Bonferroni - Model 56 ROG_SIM_DATA';
title2 ' ' ;run;

/**** RAW *****/

data model56 (keep=sim hyp raw_p);
 set  posthoc.results791 posthoc.results977 ;
  run;proc sort;by sim hyp;
data model56;  set model56;    n=_n_;     run;
proc transpose out=first(drop=_name_); by n sim; var hyp raw_p ; run;
proc transpose data=first out=posthoc.m56raw(drop=_name_) prefix=praw;
by sim;   var col1;   proc print;
title1 'RAW - Model 56 ROG_SIM_DATA';
title2 ' ' ;run;

data posthoc.m56raw (rename=(praw2=prawN));
 set posthoc.m56raw;run;
 data posthoc.m56raw (drop=praw1 );
 set posthoc.m56raw;
model=56;run;
 proc print;
title1 ' RAW - Model 56 ROG_SIM_DATA';
title2 ' ' ;run;
/*** END RAW ****/
/**** STEP BONFERRONI ****/

data model56 (keep=sim hyp stpbon_p);
 set  posthoc.results791 posthoc.results977 ;
  run;proc sort;by sim hyp;
data model56;  set model56;    n=_n_;     run;
proc transpose out=first(drop=_name_); by n sim; var hyp stpbon_p ;
run;
proc transpose data=first out=posthoc.m56stpBon(drop=_name_)
prefix=stpbon;
by sim;   var col1;   proc print;
title1 'S Bonferroni - Model 56 ROG_SIM_DATA';
title2 ' ' ;run;

data posthoc.m56stpbon (rename=(stpbon2=stpbonN));
```

```
 set posthoc.m56stpbon;run;
 data posthoc.m56stpbon (drop=stpbon1 );
 set posthoc.m56stpbon;
model=56;run;
 proc print;
title1 'S Bonferroni - Model 56 ROG_SIM_DATA';
title2 ' ' ;run;
/*** END STEP BON ******/
/*** SIDAK****/

data model56 (keep=sim hyp sid_p);
 set  posthoc.results791 posthoc.results977 ;
  run;proc sort;by sim hyp;
data model56;  set model56;    n=_n_;     run;
proc transpose out=first(drop=_name_); by n sim; var hyp sid_p ; run;
proc transpose data=first out=posthoc.m56sid(drop=_name_) prefix=psid;
by sim;   var col1;   proc print;
title1 'S - Model 56 ROG_SIM_DATA';
title2 ' ' ;run;

data posthoc.m56sid (rename=(psid2=psidN));
 set posthoc.m56sid;run;
 data posthoc.m56sid (drop=psid1 );
 set posthoc.m56sid;
model=56;run;
 proc print;
title1 ' S - Model 56 ROG_SIM_DATA';
title2 ' ' ;run;
/**** END SIDAK ***/
/*** STEP SIDAK****/

data model56 (keep=sim hyp stpsid_p);
 set  posthoc.results791 posthoc.results977 ;
  run;proc sort;by sim hyp;
data model56;  set model56;    n=_n_;     run;
proc transpose out=first(drop=_name_); by n sim; var hyp stpsid_p ;
run;
proc transpose data=first out=posthoc.m56stpsid(drop=_name_)
prefix=pstpsid;
by sim;   var col1;   proc print;
title1 'SS - Model 56 ROG_SIM_DATA';
title2 ' ' ;run;

data posthoc.m56stpsid (rename=(pstpsid2=pstpsidN));
 set posthoc.m56stpsid;run;
 data posthoc.m56stpsid (drop=pstpsid1 );
 set posthoc.m56stpsid;
model=56;run;
 proc print;
title1 ' SS - Model 56 ROG_SIM_DATA';
title2 ' ' ;run;
/*** END STEP SIDAK *****/
```

Appendix H
SAS Commands for Holm's Step-Down Procedure

/*Holm's Step Down Procedure*/

```
/*Marginal Homogeneity*/
data pvalsMO;
input comparison$ raw_p;
cards;
dm34  0.097
dm45  0.178
dm14  0.178
dm46  0.327
dm23  0.341
dm36  0.341
dm24  0.341
dm13  0.571
dm35  0.571
dm12  0.657
dm16  0.657
dm25  0.657
dm56  0.657
dm15  1.00
dm26  1.00
;
proc multtest pdata=pvalsMO holm;
title ‘MCP Procedure: Marginal Homogeneity';
run;

/*Marginal Heterogeneity*/
data pvalsMG;
input comparison$ raw_p;
cards;
dm46  0.041
dm24  0.052
dm45  0.125
dm14  0.146
dm36  0.346
dm23  0.362
dm34  0.444
dm35  0.570
dm16  0.584
dm12  0.610
dm13  0.641
dm56  0.667
dm25  0.695
dm15  0.908
dm26  0.969
;
proc multtest pdata=pvalsMG holm;
title ‘MCP Procedure: Marginal Heterogeneity';
run;
```

Appendix I
SAS Output From Commands for Holm’s Step-Down Procedure for Table 14

```
                     MCP Procedure: Marginal Homogeneity


                           The Multtest Procedure

                                  p-Values

                                          Stepdown                  Stepdown
Test         Raw    Bonferroni    Bonferroni         Sidak         Sidak

   1      0.0970        1.0000        1.0000        0.7836        0.7836
   2      0.1780        1.0000        1.0000        0.9471        0.9357
   3      0.1780        1.0000        1.0000        0.9471        0.9357
   4      0.3270        1.0000        1.0000        0.9974        0.9914
   5      0.3410        1.0000        1.0000        0.9981        0.9914
   6      0.3410        1.0000        1.0000        0.9981        0.9914
   7      0.3410        1.0000        1.0000        0.9981        0.9914
   8      0.5710        1.0000        1.0000        1.0000        0.9989
   9      0.5710        1.0000        1.0000        1.0000        0.9989
  10      0.6570        1.0000        1.0000        1.0000        0.9989
  11      0.6570        1.0000        1.0000        1.0000        0.9989
  12      0.6570        1.0000        1.0000        1.0000        0.9989
  13      0.6570        1.0000        1.0000        1.0000        0.9989
  14      1.0000        1.0000        1.0000        1.0000        1.0000
  15      1.0000        1.0000        1.0000        1.0000        1.0000
```

```
                    MCP Procedure: Marginal Heterogeneity


                             The Multtest Procedure

                                    p-Values

                                               Stepdown                Stepdown
Test          Raw    Bonferroni    Bonferroni         Sidak           Sidak

   1       0.0410        0.6150        0.6150        0.4663          0.4663
   2       0.0520        0.7800        0.7280        0.5511          0.5265
   3       0.1250        1.0000        1.0000        0.8651          0.8238
   4       0.1460        1.0000        1.0000        0.9063          0.8495
   5       0.3460        1.0000        1.0000        0.9983          0.9906
   6       0.3620        1.0000        1.0000        0.9988          0.9906
   7       0.4440        1.0000        1.0000        0.9998          0.9949
   8       0.5700        1.0000        1.0000        1.0000          0.9988
   9       0.5840        1.0000        1.0000        1.0000          0.9988
  10       0.6100        1.0000        1.0000        1.0000          0.9988
  11       0.6410        1.0000        1.0000        1.0000          0.9988
  12       0.6670        1.0000        1.0000        1.0000          0.9988
  13       0.6950        1.0000        1.0000        1.0000          0.9988
  14       0.9080        1.0000        1.0000        1.0000          0.9988
  15       0.9690        1.0000        1.0000        1.0000          0.9988
```

Appendix J
SAS Commands Summarizing MCP Results by the Number of Possible Pair-wise Comparisons

Because each of the 64 possible GHeP models allows for all or a subset of the 15 possible pair-wise comparisons, 15 variables per multiple comparison procedure were created. For example, variables **pbonA, pbonB, …. pbonP** were created to represent the Bonferroni adjusted p-value from each of the fifteen pair-wise comparisons. The suffix 'A' through 'P' (skipping 'O') uniquely represents what two heterogeneous partial agreement parameters are involved in the pair-wise comparison. Variable **pbonA** represents the pair-wise comparison of d_5^1 and d_5^2, **pbonB** represents the pair-wise comparison of d_5^1 and d_5^3, …, and **pbonP** represents the pair-wise comparison of d_5^5 and d_5^6. Similar variables were created for unadjusted p-values and the Holm's -Bonferroni, Sidak, and Holm's-Sidak adjusted p-values.

```
libname posthoc
"C:\aaPhDSimulations\HomoG\GhepAtyp4a\multtest_results\pvalues";


data posthoc.QHomGraw_Data15;
 set posthoc.m1raw;


if (.<prawA<=.05) then c12=1; if prawA > 0.05 then c12=0;
if (.<prawB<=.05) then c13=1; if prawB > 0.05 then c13=0;
if (.<prawC<=.05) then c14=1; if prawC > 0.05 then c14=0;
if (.<prawD<=.05) then c15=1; if prawD > 0.05 then c15=0;
if (.<prawE<=.05) then c16=1; if prawE > 0.05 then c16=0;

if (.<prawF<=.05) then c23=1; if prawF > 0.05 then c23=0;
if (.<prawG<=.05) then c24=1; if prawG > 0.05 then c24=0;
if (.<prawH<=.05) then c25=1; if prawH > 0.05 then c25=0;
if (.<prawI<=.05) then c26=1; if prawI > 0.05 then c26=0;

if (.<prawJ<=.05) then c34=1; if prawJ > 0.05 then c34=0;
if (.<prawK<=.05) then c35=1; if prawK > 0.05 then c35=0;
if (.<prawL<=.05) then c36=1; if prawL > 0.05 then c36=0;

if (.<prawM<=.05) then c45=1; if prawM > 0.05 then c45=0;
if (.<prawN<=.05) then c46=1; if prawN > 0.05 then c46=0;

if (.<prawP<=.05) then c56=1; if prawP > 0.05 then c56=0;

Rater1=Sum(c12,c13,c14,c15,c16);
Rater2=Sum(c12,c23,c24,c25,c26);
Rater3=Sum(c13,c23,c34,c35,c36);
Rater4=Sum(c14,c24,c34,c45,c46);
Rater5=Sum(c15,c25,c35,c45,c56);
Rater6=Sum(c16,c26,c36,c46,c56);


R1not4=Sum(c12,c13,c15,c16);
R2not4=Sum(c12,c23,c25,c26);
R3not4=Sum(c13,c23,c35,c36);
Rater4=Sum(c14,c24,c34,c45,c46);
R5not4=Sum(c15,c25,c35,c56);
R6not4=Sum(c16,c26,c36,c56);run;
```

```
proc freq;tables model;
title1 '15 Homog_GSimulated Data - 1K';
title2 'Results of RAW MCPs';run;
proc freq;tables c12 c13 c14 c15 c16 c23 c24 c25 c26 c34 c35 c36 c45
c46 c56;
title1 '15 Homog_GSimulated Data - 1K';
title2 'Results of RAW MCPs';run;
proc freq;tables Rater1 Rater2 Rater3 Rater4 Rater5 Rater6;
title1 '15 Homog_GSimulated Data - 1K';
title2 'Results of RAW MCPs';run;
proc freq;tables R1not4 R2not4 R3not4 Rater4 R5not4 R6not4;
run;
/*10 P-W Comparisons*/
data posthoc.QHomGraw_Data10;
 set posthoc.m2raw posthoc.m3raw posthoc.m4raw posthoc.m5raw
posthoc.m6raw posthoc.m7raw;

if (.<prawA<=.05) then c12=1; if prawA > 0.05 then c12=0;
if (.<prawB<=.05) then c13=1; if prawB > 0.05 then c13=0;
if (.<prawC<=.05) then c14=1; if prawC > 0.05 then c14=0;
if (.<prawD<=.05) then c15=1; if prawD > 0.05 then c15=0;
if (.<prawE<=.05) then c16=1; if prawE > 0.05 then c16=0;

if (.<prawF<=.05) then c23=1; if prawF > 0.05 then c23=0;
if (.<prawG<=.05) then c24=1; if prawG > 0.05 then c24=0;
if (.<prawH<=.05) then c25=1; if prawH > 0.05 then c25=0;
if (.<prawI<=.05) then c26=1; if prawI > 0.05 then c26=0;

if (.<prawJ<=.05) then c34=1; if prawJ > 0.05 then c34=0;
if (.<prawK<=.05) then c35=1; if prawK > 0.05 then c35=0;
if (.<prawL<=.05) then c36=1; if prawL > 0.05 then c36=0;

if (.<prawM<=.05) then c45=1; if prawM > 0.05 then c45=0;
if (.<prawN<=.05) then c46=1; if prawN > 0.05 then c46=0;

if (.<prawP<=.05) then c56=1; if prawP > 0.05 then c56=0;

Rater1=Sum(c12,c13,c14,c15,c16);
Rater2=Sum(c12,c23,c24,c25,c26);
Rater3=Sum(c13,c23,c34,c35,c36);
Rater4=Sum(c14,c24,c34,c45,c46);
Rater5=Sum(c15,c25,c35,c45,c56);
Rater6=Sum(c16,c26,c36,c46,c56);


R1not4=Sum(c12,c13,c15,c16);
R2not4=Sum(c12,c23,c25,c26);
R3not4=Sum(c13,c23,c35,c36);
Rater4=Sum(c14,c24,c34,c45,c46);
R5not4=Sum(c15,c25,c35,c56);
R6not4=Sum(c16,c26,c36,c56);run;

proc freq;tables model;
title1 '10 Homog_GSimulated Data - 1K';
title2 'Results of RAW MCPs';run;
```

```
proc freq;tables c12 c13 c14 c15 c16 c23 c24 c25 c26 c34 c35 c36 c45
c46 c56;
title1 '10 Homog_GSimulated Data - 1K';
title2 'Results of RAW MCPs';run;
proc freq;tables Rater1 Rater2 Rater3 Rater4 Rater5 Rater6;
title1 '10 Homog_GSimulated Data - 1K';
title2 'Results of RAW MCPs';run;
proc freq;tables R1not4 R2not4 R3not4 Rater4 R5not4 R6not4;
run;

data posthoc.HomG_Data6;
 set posthoc.m8raw posthoc.m9raw posthoc.m10raw
posthoc.m11raw posthoc.m12raw posthoc.m13raw
posthoc.m15raw posthoc.m16raw
posthoc.m17raw posthoc.m18raw posthoc.m19raw
posthoc.m22raw  ;


if (.<prawA<=.05) then c12=1; if prawA > 0.05 then c12=0;
if (.<prawB<=.05) then c13=1; if prawB > 0.05 then c13=0;
if (.<prawC<=.05) then c14=1; if prawC > 0.05 then c14=0;
if (.<prawD<=.05) then c15=1; if prawD > 0.05 then c15=0;
if (.<prawE<=.05) then c16=1; if prawE > 0.05 then c16=0;

if (.<prawF<=.05) then c23=1; if prawF > 0.05 then c23=0;
if (.<prawG<=.05) then c24=1; if prawG > 0.05 then c24=0;
if (.<prawH<=.05) then c25=1; if prawH > 0.05 then c25=0;
if (.<prawI<=.05) then c26=1; if prawI > 0.05 then c26=0;

if (.<prawJ<=.05) then c34=1; if prawJ > 0.05 then c34=0;
if (.<prawK<=.05) then c35=1; if prawK > 0.05 then c35=0;
if (.<prawL<=.05) then c36=1; if prawL > 0.05 then c36=0;

if (.<prawM<=.05) then c45=1; if prawM > 0.05 then c45=0;
if (.<prawN<=.05) then c46=1; if prawN > 0.05 then c46=0;

if (.<prawP<=.05) then c56=1; if prawP > 0.05 then c56=0;

Rater1=Sum(c12,c13,c14,c15,c16);
Rater2=Sum(c12,c23,c24,c25,c26);
Rater3=Sum(c13,c23,c34,c35,c36);
Rater4=Sum(c14,c24,c34,c45,c46);
Rater5=Sum(c15,c25,c35,c45,c56);
Rater6=Sum(c16,c26,c36,c46,c56);


R1not4=Sum(c12,c13,c15,c16);
R2not4=Sum(c12,c23,c25,c26);
R3not4=Sum(c13,c23,c35,c36);
Rater4=Sum(c14,c24,c34,c45,c46);
R5not4=Sum(c15,c25,c35,c56);
R6not4=Sum(c16,c26,c36,c56);run;

proc freq;tables model;
title1 '6 Homog_GSimulated Data - 1K';
title2 'Results of RAW MCPs';run;
```

```
proc freq;tables c12 c13 c14 c15 c16 c23 c24 c25 c26 c34 c35 c36 c45
c46 c56;
title1 '6 Homog_GSimulated Data - 1K';
title2 'Results of RAW MCPs';run;
proc freq;tables Rater1 Rater2 Rater3 Rater4 Rater5 Rater6;
title1 '6 Homog_GSimulated Data - 1K';
title2 'Results of RAW MCPs';run;
proc freq;tables R1not4 R2not4 R3not4 Rater4 R5not4 R6not4;
run;
/*THREE P-W Comparison*/
data posthoc.HomG_Data3;
 set
posthoc.m23raw posthoc.m25raw posthoc.m26raw posthoc.m28raw
posthoc.m29raw posthoc.m31raw posthoc.m33raw posthoc.m34raw
posthoc.m36raw
posthoc.m38raw
posthoc.m41raw posthoc.m42raw  ;

if (.<prawA<=.05) then c12=1; if prawA > 0.05 then c12=0;
if (.<prawB<=.05) then c13=1; if prawB > 0.05 then c13=0;
if (.<prawC<=.05) then c14=1; if prawC > 0.05 then c14=0;
if (.<prawD<=.05) then c15=1; if prawD > 0.05 then c15=0;
if (.<prawE<=.05) then c16=1; if prawE > 0.05 then c16=0;

if (.<prawF<=.05) then c23=1; if prawF > 0.05 then c23=0;
if (.<prawG<=.05) then c24=1; if prawG > 0.05 then c24=0;
if (.<prawH<=.05) then c25=1; if prawH > 0.05 then c25=0;
if (.<prawI<=.05) then c26=1; if prawI > 0.05 then c26=0;

if (.<prawJ<=.05) then c34=1; if prawJ > 0.05 then c34=0;
if (.<prawK<=.05) then c35=1; if prawK > 0.05 then c35=0;
if (.<prawL<=.05) then c36=1; if prawL > 0.05 then c36=0;

if (.<prawM<=.05) then c45=1; if prawM > 0.05 then c45=0;
if (.<prawN<=.05) then c46=1; if prawN > 0.05 then c46=0;

if (.<prawP<=.05) then c56=1; if prawP > 0.05 then c56=0;

Rater1=Sum(c12,c13,c14,c15,c16);
Rater2=Sum(c12,c23,c24,c25,c26);
Rater3=Sum(c13,c23,c34,c35,c36);
Rater4=Sum(c14,c24,c34,c45,c46);
Rater5=Sum(c15,c25,c35,c45,c56);
Rater6=Sum(c16,c26,c36,c46,c56);


R1not4=Sum(c12,c13,c15,c16);
R2not4=Sum(c12,c23,c25,c26);
R3not4=Sum(c13,c23,c35,c36);
Rater4=Sum(c14,c24,c34,c45,c46);
R5not4=Sum(c15,c25,c35,c56);
R6not4=Sum(c16,c26,c36,c56);run;


proc freq;tables model;
title1 '3 Homog_GSimulated Data - 1K';
title2 'Results of RAW MCPs';run;
```

```
proc freq;tables c12 c13 c14 c15 c16 c23 c24 c25 c26 c34 c35 c36 c45
c46 c56;
title1 '3 Homog_GSimulated Data - 1K';
title2 'Results of RAW MCPs';run;
proc freq;tables Rater1 Rater2 Rater3 Rater4 Rater5 Rater6;
title1 '3 Homog_GSimulated Data - 1K';
title2 'Results of RAW MCPs';run;
proc freq;tables R1not4 R2not4 R3not4 Rater4 R5not4 R6not4;
run;

/*ONE P-W Comparison*/
data posthoc.HomG_Data1;
 set  posthoc.m45raw
posthoc.m49raw  posthoc.m51raw posthoc.m56raw ;


if (.<prawA<=.05) then c12=1; if prawA > 0.05 then c12=0;
if (.<prawB<=.05) then c13=1; if prawB > 0.05 then c13=0;
if (.<prawC<=.05) then c14=1; if prawC > 0.05 then c14=0;
if (.<prawD<=.05) then c15=1; if prawD > 0.05 then c15=0;
if (.<prawE<=.05) then c16=1; if prawE > 0.05 then c16=0;

if (.<prawF<=.05) then c23=1; if prawF > 0.05 then c23=0;
if (.<prawG<=.05) then c24=1; if prawG > 0.05 then c24=0;
if (.<prawH<=.05) then c25=1; if prawH > 0.05 then c25=0;
if (.<prawI<=.05) then c26=1; if prawI > 0.05 then c26=0;

if (.<prawJ<=.05) then c34=1; if prawJ > 0.05 then c34=0;
if (.<prawK<=.05) then c35=1; if prawK > 0.05 then c35=0;
if (.<prawL<=.05) then c36=1; if prawL > 0.05 then c36=0;

if (.<prawM<=.05) then c45=1; if prawM > 0.05 then c45=0;
if (.<prawN<=.05) then c46=1; if prawN > 0.05 then c46=0;

if (.<prawP<=.05) then c56=1; if prawP > 0.05 then c56=0;

Rater1=Sum(c12,c13,c14,c15,c16);
Rater2=Sum(c12,c23,c24,c25,c26);
Rater3=Sum(c13,c23,c34,c35,c36);
Rater4=Sum(c14,c24,c34,c45,c46);
Rater5=Sum(c15,c25,c35,c45,c56);
Rater6=Sum(c16,c26,c36,c46,c56);


R1not4=Sum(c12,c13,c15,c16);
R2not4=Sum(c12,c23,c25,c26);
R3not4=Sum(c13,c23,c35,c36);
Rater4=Sum(c14,c24,c34,c45,c46);
R5not4=Sum(c15,c25,c35,c56);
R6not4=Sum(c16,c26,c36,c56);run;


proc freq;tables model;
title1 '1 Homog_GSimulated Data - 1K';
title2 'Results of RAW MCPs';run;
proc freq;tables c12 c13 c14 c15 c16 c23 c24 c25 c26 c34 c35 c36 c45
c46 c56;
```

```
title1 '1 Homog_GSimulated Data - 1K';
title2 'Results of RAW MCPs';run;
proc freq;tables Rater1 Rater2 Rater3 Rater4 Rater5 Rater6;
title1 '1 Homog_GSimulated Data - 1K';
title2 'Results of RAW MCPs';run;
proc freq;tables R1not4 R2not4 R3not4 Rater4 R5not4 R6not4;
run;


data posthoc.QHomGbon_Data15;
 set
posthoc.m1bon ;

if (.<pbonA<=.05) then c12=1; if pbonA > 0.05 then c12=0;
if (.<pbonB<=.05) then c13=1; if pbonB > 0.05 then c13=0;
if (.<pbonC<=.05) then c14=1; if pbonC > 0.05 then c14=0;
if (.<pbonD<=.05) then c15=1; if pbonD > 0.05 then c15=0;
if (.<pbonE<=.05) then c16=1; if pbonE > 0.05 then c16=0;

if (.<pbonF<=.05) then c23=1; if pbonF > 0.05 then c23=0;
if (.<pbonG<=.05) then c24=1; if pbonG > 0.05 then c24=0;
if (.<pbonH<=.05) then c25=1; if pbonH > 0.05 then c25=0;
if (.<pbonI<=.05) then c26=1; if pbonI > 0.05 then c26=0;

if (.<pbonJ<=.05) then c34=1; if pbonJ > 0.05 then c34=0;
if (.<pbonK<=.05) then c35=1; if pbonK > 0.05 then c35=0;
if (.<pbonL<=.05) then c36=1; if pbonL > 0.05 then c36=0;

if (.<pbonM<=.05) then c45=1; if pbonM > 0.05 then c45=0;
if (.<pbonN<=.05) then c46=1; if pbonN > 0.05 then c46=0;

if (.<pbonP<=.05) then c56=1; if pbonP > 0.05 then c56=0;

Rater1=Sum(c12,c13,c14,c15,c16);
Rater2=Sum(c12,c23,c24,c25,c26);
Rater3=Sum(c13,c23,c34,c35,c36);
Rater4=Sum(c14,c24,c34,c45,c46);
Rater5=Sum(c15,c25,c35,c45,c56);
Rater6=Sum(c16,c26,c36,c46,c56);

R1not4=Sum(c12,c13,c15,c16);
R2not4=Sum(c12,c23,c25,c26);
R3not4=Sum(c13,c23,c35,c36);
Rater4=Sum(c14,c24,c34,c45,c46);
R5not4=Sum(c15,c25,c35,c56);
R6not4=Sum(c16,c26,c36,c56);
run;


proc freq;tables model;
title1 '15 Homog_GSimulated Data - 1K';
title2 'Results of BON MCPs';run;
proc freq;tables c12 c13 c14 c15 c16 c23 c24 c25 c26 c34 c35 c36 c45
c46 c56;
title1 '15 Homog_GSimulated Data - 1K';
title2 'Results of BON MCPs';run;
proc freq;tables Rater1 Rater2 Rater3 Rater4 Rater5 Rater6;
```

```
title1 '15 Homog_GSimulated Data - 1K';
title2 'Results of BON MCPs';run;
proc freq;tables R1not4 R2not4 R3not4 Rater4 R5not4 R6not4;
run;

data posthoc.QHomGbon_Data10;
 set posthoc.m2bon posthoc.m3bon posthoc.m4bon posthoc.m5bon
posthoc.m6bon posthoc.m7bon ;

if (.<pbonA<=.05) then c12=1; if pbonA > 0.05 then c12=0;
if (.<pbonB<=.05) then c13=1; if pbonB > 0.05 then c13=0;
if (.<pbonC<=.05) then c14=1; if pbonC > 0.05 then c14=0;
if (.<pbonD<=.05) then c15=1; if pbonD > 0.05 then c15=0;
if (.<pbonE<=.05) then c16=1; if pbonE > 0.05 then c16=0;

if (.<pbonF<=.05) then c23=1; if pbonF > 0.05 then c23=0;
if (.<pbonG<=.05) then c24=1; if pbonG > 0.05 then c24=0;
if (.<pbonH<=.05) then c25=1; if pbonH > 0.05 then c25=0;
if (.<pbonI<=.05) then c26=1; if pbonI > 0.05 then c26=0;

if (.<pbonJ<=.05) then c34=1; if pbonJ > 0.05 then c34=0;
if (.<pbonK<=.05) then c35=1; if pbonK > 0.05 then c35=0;
if (.<pbonL<=.05) then c36=1; if pbonL > 0.05 then c36=0;

if (.<pbonM<=.05) then c45=1; if pbonM > 0.05 then c45=0;
if (.<pbonN<=.05) then c46=1; if pbonN > 0.05 then c46=0;

if (.<pbonP<=.05) then c56=1; if pbonP > 0.05 then c56=0;

Rater1=Sum(c12,c13,c14,c15,c16);
Rater2=Sum(c12,c23,c24,c25,c26);
Rater3=Sum(c13,c23,c34,c35,c36);
Rater4=Sum(c14,c24,c34,c45,c46);
Rater5=Sum(c15,c25,c35,c45,c56);
Rater6=Sum(c16,c26,c36,c46,c56);

R1not4=Sum(c12,c13,c15,c16);
R2not4=Sum(c12,c23,c25,c26);
R3not4=Sum(c13,c23,c35,c36);
Rater4=Sum(c14,c24,c34,c45,c46);
R5not4=Sum(c15,c25,c35,c56);
R6not4=Sum(c16,c26,c36,c56);
run;


proc freq;tables model;
title1 '10 Homog_GSimulated Data - 1K';
title2 'Results of BON MCPs';run;
proc freq;tables c12 c13 c14 c15 c16 c23 c24 c25 c26 c34 c35 c36 c45
c46 c56;
title1 '10 Homog_GSimulated Data - 1K';
title2 'Results of BON MCPs';run;
proc freq;tables Rater1 Rater2 Rater3 Rater4 Rater5 Rater6;
title1 '10 Homog_GSimulated Data - 1K';
title2 'Results of BON MCPs';run;
proc freq;tables R1not4 R2not4 R3not4 Rater4 R5not4 R6not4;
run;
```

```
data posthoc.QHomGbon_Data6;
 set posthoc.m8bon posthoc.m9bon posthoc.m10bon
posthoc.m11bon posthoc.m12bon posthoc.m13bon
posthoc.m15bon posthoc.m16bon
posthoc.m17bon posthoc.m18bon posthoc.m19bon posthoc.m22bon;

if (.<pbonA<=.05) then c12=1; if pbonA > 0.05 then c12=0;
if (.<pbonB<=.05) then c13=1; if pbonB > 0.05 then c13=0;
if (.<pbonC<=.05) then c14=1; if pbonC > 0.05 then c14=0;
if (.<pbonD<=.05) then c15=1; if pbonD > 0.05 then c15=0;
if (.<pbonE<=.05) then c16=1; if pbonE > 0.05 then c16=0;

if (.<pbonF<=.05) then c23=1; if pbonF > 0.05 then c23=0;
if (.<pbonG<=.05) then c24=1; if pbonG > 0.05 then c24=0;
if (.<pbonH<=.05) then c25=1; if pbonH > 0.05 then c25=0;
if (.<pbonI<=.05) then c26=1; if pbonI > 0.05 then c26=0;

if (.<pbonJ<=.05) then c34=1; if pbonJ > 0.05 then c34=0;
if (.<pbonK<=.05) then c35=1; if pbonK > 0.05 then c35=0;
if (.<pbonL<=.05) then c36=1; if pbonL > 0.05 then c36=0;

if (.<pbonM<=.05) then c45=1; if pbonM > 0.05 then c45=0;
if (.<pbonN<=.05) then c46=1; if pbonN > 0.05 then c46=0;

if (.<pbonP<=.05) then c56=1; if pbonP > 0.05 then c56=0;

Rater1=Sum(c12,c13,c14,c15,c16);
Rater2=Sum(c12,c23,c24,c25,c26);
Rater3=Sum(c13,c23,c34,c35,c36);
Rater4=Sum(c14,c24,c34,c45,c46);
Rater5=Sum(c15,c25,c35,c45,c56);
Rater6=Sum(c16,c26,c36,c46,c56);

R1not4=Sum(c12,c13,c15,c16);
R2not4=Sum(c12,c23,c25,c26);
R3not4=Sum(c13,c23,c35,c36);
Rater4=Sum(c14,c24,c34,c45,c46);
R5not4=Sum(c15,c25,c35,c56);
R6not4=Sum(c16,c26,c36,c56);
run;


proc freq;tables model;
title1 '6 Homog_GSimulated Data - 1K';
title2 'Results of BON MCPs';run;
proc freq;tables c12 c13 c14 c15 c16 c23 c24 c25 c26 c34 c35 c36 c45
c46 c56;
title1 '6 Homog_GSimulated Data - 1K';
title2 'Results of BON MCPs';run;
proc freq;tables Rater1 Rater2 Rater3 Rater4 Rater5 Rater6;
title1 '6 Homog_GSimulated Data - 1K';
title2 'Results of BON MCPs';run;
proc freq;tables R1not4 R2not4 R3not4 Rater4 R5not4 R6not4;
run;
```

```
data posthoc.QHomGbon_Data3;
 set posthoc.m23bon posthoc.m25bon posthoc.m26bon posthoc.m28bon
posthoc.m29bon posthoc.m31bon posthoc.m33bon posthoc.m34bon
posthoc.m36bon  posthoc.m38bon
posthoc.m41bon posthoc.m42bon ;


if (.<pbonA<=.05) then c12=1; if pbonA > 0.05 then c12=0;
if (.<pbonB<=.05) then c13=1; if pbonB > 0.05 then c13=0;
if (.<pbonC<=.05) then c14=1; if pbonC > 0.05 then c14=0;
if (.<pbonD<=.05) then c15=1; if pbonD > 0.05 then c15=0;
if (.<pbonE<=.05) then c16=1; if pbonE > 0.05 then c16=0;

if (.<pbonF<=.05) then c23=1; if pbonF > 0.05 then c23=0;
if (.<pbonG<=.05) then c24=1; if pbonG > 0.05 then c24=0;
if (.<pbonH<=.05) then c25=1; if pbonH > 0.05 then c25=0;
if (.<pbonI<=.05) then c26=1; if pbonI > 0.05 then c26=0;

if (.<pbonJ<=.05) then c34=1; if pbonJ > 0.05 then c34=0;
if (.<pbonK<=.05) then c35=1; if pbonK > 0.05 then c35=0;
if (.<pbonL<=.05) then c36=1; if pbonL > 0.05 then c36=0;

if (.<pbonM<=.05) then c45=1; if pbonM > 0.05 then c45=0;
if (.<pbonN<=.05) then c46=1; if pbonN > 0.05 then c46=0;

if (.<pbonP<=.05) then c56=1; if pbonP > 0.05 then c56=0;

Rater1=Sum(c12,c13,c14,c15,c16);
Rater2=Sum(c12,c23,c24,c25,c26);
Rater3=Sum(c13,c23,c34,c35,c36);
Rater4=Sum(c14,c24,c34,c45,c46);
Rater5=Sum(c15,c25,c35,c45,c56);
Rater6=Sum(c16,c26,c36,c46,c56);

R1not4=Sum(c12,c13,c15,c16);
R2not4=Sum(c12,c23,c25,c26);
R3not4=Sum(c13,c23,c35,c36);
Rater4=Sum(c14,c24,c34,c45,c46);
R5not4=Sum(c15,c25,c35,c56);
R6not4=Sum(c16,c26,c36,c56);
run;

proc freq;tables model;
title1 '3 Homog_GSimulated Data - 1K';
title2 'Results of BON MCPs';run;
proc freq;tables c12 c13 c14 c15 c16 c23 c24 c25 c26 c34 c35 c36 c45
c46 c56;
title1 '3 Homog_GSimulated Data - 1K';
title2 'Results of BON MCPs';run;
proc freq;tables Rater1 Rater2 Rater3 Rater4 Rater5 Rater6;
title1 '3 Homog_GSimulated Data - 1K';
title2 'Results of BON MCPs';run;
proc freq;tables R1not4 R2not4 R3not4 Rater4 R5not4 R6not4;
run;
```

```
data posthoc.QHomGbon_Data1;
 set posthoc.m45bon  posthoc.m49bon  posthoc.m51bon posthoc.m56bon ;


if (.<pbonA<=.05) then c12=1; if pbonA > 0.05 then c12=0;
if (.<pbonB<=.05) then c13=1; if pbonB > 0.05 then c13=0;
if (.<pbonC<=.05) then c14=1; if pbonC > 0.05 then c14=0;
if (.<pbonD<=.05) then c15=1; if pbonD > 0.05 then c15=0;
if (.<pbonE<=.05) then c16=1; if pbonE > 0.05 then c16=0;

if (.<pbonF<=.05) then c23=1; if pbonF > 0.05 then c23=0;
if (.<pbonG<=.05) then c24=1; if pbonG > 0.05 then c24=0;
if (.<pbonH<=.05) then c25=1; if pbonH > 0.05 then c25=0;
if (.<pbonI<=.05) then c26=1; if pbonI > 0.05 then c26=0;

if (.<pbonJ<=.05) then c34=1; if pbonJ > 0.05 then c34=0;
if (.<pbonK<=.05) then c35=1; if pbonK > 0.05 then c35=0;
if (.<pbonL<=.05) then c36=1; if pbonL > 0.05 then c36=0;

if (.<pbonM<=.05) then c45=1; if pbonM > 0.05 then c45=0;
if (.<pbonN<=.05) then c46=1; if pbonN > 0.05 then c46=0;

if (.<pbonP<=.05) then c56=1; if pbonP > 0.05 then c56=0;

Rater1=Sum(c12,c13,c14,c15,c16);
Rater2=Sum(c12,c23,c24,c25,c26);
Rater3=Sum(c13,c23,c34,c35,c36);
Rater4=Sum(c14,c24,c34,c45,c46);
Rater5=Sum(c15,c25,c35,c45,c56);
Rater6=Sum(c16,c26,c36,c46,c56);

R1not4=Sum(c12,c13,c15,c16);
R2not4=Sum(c12,c23,c25,c26);
R3not4=Sum(c13,c23,c35,c36);
Rater4=Sum(c14,c24,c34,c45,c46);
R5not4=Sum(c15,c25,c35,c56);
R6not4=Sum(c16,c26,c36,c56);
run;


proc freq;tables model;
title1 '1 Homog_GSimulated Data - 1K';
title2 'Results of BON MCPs';run;
proc freq;tables c12 c13 c14 c15 c16 c23 c24 c25 c26 c34 c35 c36 c45
c46 c56;
title1 '1 Homog_GSimulated Data - 1K';
title2 'Results of BON MCPs';run;
proc freq;tables Rater1 Rater2 Rater3 Rater4 Rater5 Rater6;
title1 '1 Homog_GSimulated Data - 1K';
title2 'Results of BON MCPs';run;
proc freq;tables R1not4 R2not4 R3not4 Rater4 R5not4 R6not4;run;
```

BIBLIOGRAPHY

Agresti, A. *Categorical data analysis*, Wiley-Interscience, New York, (2002).

Akaike, H. (1974). A new look at statistical model identification. IEEE Transactions on Automatic Control, 19: 716-722.

Bishop, YMM, Fienberg, SE, and Holland, PW. *Discrete multivariate analysis: theory and practice,* MIT Press, Cambridge, MA (1975).

Belsley, DA, Kuh, E, and Welsch, RE. *Regression diagnostics: identifying influential data and sources of collinearity.* Wiley, New York (1980).

Burnham, K.P., & Anderson, D.R. 1998. *Model selection and inference.* Springer.

Christensen, R. *Log-linear models and logistic regression,* Springer, New York (1997).

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-39 (1960).

Fleiss, JL, Cohen, J, and Everitt, BS. Large sample standard errors of kappa and weighted kappa. Psychological Bulletin, 72, 323-327 (1969).

Fleiss, JL. *Statistical Methods for Rates and Proportions,* Wiley, New York (1981).

Goodman, LA (1968) The Analysis of Cross-Classified Data: Independence, Quasi-Independence and Interactions in Contingency Tables with or Without Missing Entries. *Journal of the American Statistical Association*, 63(324):1091-1131.

Hochberg, Y and Tamhane, AC. *Multiple Comparison Procedures*. John Wiley & Sons, Inc., New York (1987).

Holms, S. A simple sequentially rejective multiple test procedure. *Scandanavia Journal of Statisitics*. 6:65-70 (1979).

Kennedy, WJ and Gentle, JE. *Statistical Computing*. Marcel Dekker, Inc., New York (1980).

Klockars, AJ, and Sax, G. *Multiple Comparisons*. Sage, Newbury Park, CA (1986).

Ludbrook, J. Multiple comparison procedures updated. Clinical and Experimental Pharmacology and Physiology, 25(12):1032-1037 (1998).

Marcus, R, Peritz, E, and Gabriel, KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika.* 63:655-660 (1976).

Myers, RH. *Classical and Modern Regression With Applications*. Duxbury Press, California (1990).

Rogel, A, Boelle PY and Mary, JY. Global and partial agreement among several observers. Statist. Med., 17, 489-501 (1998).

Rogel, A, and Mary, JY. 'Chance corrected agreement among observers with log-linear models' *in* Forcina, A., Hatzinger, R., Machetti, GM, Galmacci, G. (eds), *Statistical Modelling*, Proceedings for the 11th International Workshop on Statistical Modelling, Graphos, Citta di Castello, Italy, 324-331 (1996).

Rosner, B. *Fundamentals of Biostatistics*. Duxbury Press, California (1995).

Sidak, Z. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*. 62(318): 626-633 (1967).

Shaffer Popper, J. Modified sequentially rejective multiple test procedures. Journal of the American Statistical Association, 81(395), 826-831. (1986).

Shaffer Popper, J. Multiple Hypothesis Testing. Annual Reviews, 46, 561-584. (1995).

Tanner, MA and Young, MA. Modelling agreement among raters. Journal of the American Statistical Association, 80, 175-180 (1985).

Theodossi A, Spiegelhalter DJ, Jass J, Firth J, Dixon M, Leader M, Levison DA, Lindley R, Filip I, Price A, Shephard NA, Thomas S, and Thompson H (1994). Observer variation and discriminatory value of biopsy features in inflammatory bowel disease. *Gut* 35, 961-968.