AN APPLICATION OF WEIGHTED LEAST SQUARES METHODS TO THE ANALYSIS OF MEASUREMENT PROCESS COMPONENTS OF VARIABILITY IN AN OBSERVATIONAL STUDY

Julia MacMillan[1], Caroline Becker[1], Gary G. Koch[1], Maura Stokes[1], and H. Mac Vandiviere[2]
University of North Carolina, Chapel Hill, NC[1]
University of Kentucky, Lexington, KY[2]

ABSTRACT

Several statistical strategies can be applied in the assessment of the extent of observer agreement for health status data resulting from a hierarchical measurement process. Specifically, kappa-type measures of agreement can be used to investigate the extent of agreement for two determinations at two time points for each of the diagnostic procedures applied to a single subject. Interest may also lie in the extent of agreement between diagnostic procedures as it varies between two time points, and the extent of agreement of the two time points for each individual procedure. Weighted least squares methods are appropriate for characterizing the variation in those measures of agreement for sources of variation of interest, including those defined by time point as well as those defined by procedure.

This paper concerns itself with such analyses for data collected in a randomized trial involving subjects from several sites.

1. Introduction

A standard and an experimental procedure were performed on each subject simultaneously. There were seven experimental procedures, corresponding to the different types of devices under study. At two follow-up visits, referred to as Time 1 and Time 2 here, each of two observers classified the standard and test results as either positive or negative. Restated in the context of this design, the questions of interest here are:

1. Is the pattern of agreement different from one time point to another?
2. Does the pattern of agreement vary from one device to another?
3. To what extent do the two observers agree in the assessment of the tests?

EXAMPLE 1

The subjects are classified as negative or positive for the standard and experimental procedures at each time point. The determination of positive or negative reading is based on the average of two observers' readings. There is agreement if the standard and test device readings are both positive or both negative for a particular reading. In this section, these two types of agreement are grouped together as 'agreement'. Disagreement, however, could be in one of two categories:

1. Reading of the standard is classified as positive, but the test device reading is 'contra-positive' (i.e., negative).
2. Reading of the standard is negative, but the reading of the test device is contra-negative (i.e., positive).

The following 2 × 2 table illustrates the possible outcomes for the subjects at one time point:

		STANDARD REACTION −	STANDARD REACTION +
TEST REACTION	−		
	+		

There are sixteen possible profiles if one considers the responses for both time points, since the classification scheme now depends on four binary response variables. Accordingly, the subjects for each procedure are classified into one of the cells of a four by four table. Table 1 presents the 4 × 4 frequency table for subjects having the duplicate standard as the test procedure. Similar tables were constructed for the other test procedures. The vector consisting of the proportions in each row of the 4 × 4 table can be manipulated through the use of certain matrix operations and log and exponential transformations to produce functions of the proportions; these are directed at the relationships of time and observer agreement in which we are interested. Test statistics can then be constructed to investigate hypotheses concerning these functions and corresponding model parameters estimated through the use of weighted least squares (WLS) computations in accordance with the Grizzle, Starmer and Koch (GSK, 1969) methodology.

Let $\underset{\sim}{p} = (p_{11}, p_{12}, \cdots, p_{44})$ be the vector of proportions obtained from the table where $p_{jj'} = (n_{jj'}/n)$ is the proportion of subjects in the j-th row and j'-th column. The desired functions of this vector are four: the contra-negative proportion for Time 1, the contra-positive for Time 1, the contra-negative for Time 2 and the contra-positive for Time 2. The function vector is written as $F(\underset{\sim}{p})$ and calculated as follows:

$$F(\underset{\sim}{p}) = \exp(\underset{\sim}{A}_2\{\log(\underset{\sim}{A}_1 \cdot \underset{\sim}{p})\})$$

where

$$\underset{\sim}{A}_1 = \begin{bmatrix} 1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1 \\ 1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0 \\ 0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1 \\ 0&0&0&0&1&1&1&1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0 \\ 0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0 \\ 0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0 \end{bmatrix}$$

and

$$\underset{\sim}{A}_2 = \begin{bmatrix} -1&0&0&0&1&0&0&0 \\ 0&-1&0&0&0&1&0&0 \\ 0&0&-1&0&0&0&1&0 \\ 0&0&0&-1&0&0&0&1 \end{bmatrix}$$

The linearized Taylor series-based estimate of the covariance matrix of any function $F(\underset{\sim}{p}) = T_i(\underset{\sim}{p})$ is denoted by $V(T_i(\underset{\sim}{p}))$ (where $T_i(\underset{\sim}{p})$ is a transformation of $\underset{\sim}{p}$; i = 1 for a linear transformation, i = 2 for a logarithmic transformation, and i = 3 for an exponential transformation) and is

calculated as: $\underset{\sim}{H}_i \underset{\sim}{V}(\underset{\sim}{p}) \underset{\sim}{H}_i'$, where $\underset{\sim}{H}_1 = \underset{\sim}{A}_1$, $\underset{\sim}{H}_2 = \underset{\sim}{D}_{\underset{\sim}{p}}^{-1}$, $\underset{\sim}{H}_3 = \underset{\sim}{D}_{\exp(\underset{\sim}{p})}$, and $\underset{\sim}{D}_{\underset{\sim}{y}}$ is a diagonal matrix with elements of the vector $\underset{\sim}{y}$ on the diagonal. $\underset{\sim}{V}(\underset{\sim}{p})$ is often a block diagonal covariance matrix based on the product multinomial distribution. The variance of a compounded function such as $\underset{\sim}{F}(\underset{\sim}{p})$ above is obtained through the application of the chain rule for matrix differentiation.

The weighted least squares method is used to fit a model $\underset{\sim}{F}(\underset{\sim}{p}) \hat{=} \underset{\sim\sim}{X}\underset{\sim}{b}$, where $\underset{\sim}{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$. $\underset{\sim}{b}$ is estimated by $\underset{\sim}{b} = (\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{F}$, which has estimated variance $\underset{\sim}{V}_{\underset{\sim}{b}} = (\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})^{-1}$, and the goodness of fit chi-square statistic can be written as $Q = (\underset{\sim}{F} - \underset{\sim\sim}{X}\underset{\sim}{b})'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}(\underset{\sim}{F} - \underset{\sim\sim}{X}\underset{\sim}{b})$, and has degrees of freedom equal to the number of rows of $\underset{\sim}{X}$ minus the number of columns of $\underset{\sim}{X}$. General hypotheses of the form H_0: $\underset{\sim\sim}{C}\underset{\sim}{b} \hat{=} \underset{\sim}{0}$ can be tested with the Wald statistic

$$Q_C = \underset{\sim}{b}'\underset{\sim}{C}'(\underset{\sim\sim}{C}\underset{\sim}{V}_{\underset{\sim}{b}}\underset{\sim}{C}')^{-1}\underset{\sim\sim}{C}\underset{\sim}{b},$$

which has degrees of freedom equal to the rank of $\underset{\sim}{C}$.

The parameters b_1 and b_2 can be interpreted as the contra-positive and contra-negative proportions for Time 1, while b_3 and b_4 represent the difference between Time 2 and Time 1 for the contra-positive and contra-negative proportions, respectively. Table 2 includes the parameters b_3 and b_4 for each of the seven test procedures.

It is of interest to investigate whether the time differences represented by b_3 and b_4 are procedure-dependent or are uniform across procedures. In order to examine this, a new function vector $\underset{\sim}{E}(\underset{\sim}{p})$ was constructed, consisting of the b_3 and b_4 estimates for each of the seven separate procedures. The 28 × 28 covariance matrix for $\underset{\sim}{E}(\underset{\sim}{p})$ consists of the block diagonal matrix whose blocks are the seven covariance matrices for the individual procedure parameters b_3 and b_4. (These would be the lower right-hand corner of the covariance matrices calculated from the individual weighted least squares analyses detailed above.) The model $\underset{\sim}{F}(\underset{\sim}{p}) \hat{=} \underset{\sim\sim}{X}\underset{\sim}{b}$ was fitted where the design matrix was the identity matrix. The hypothesis tested was that the estimates for the time differences for the contra-negatives were equal across devices; the same hypothesis was tested for the contra-positives. Below are the contrast matrices required.

$$\underset{\sim}{C}_1 = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \end{bmatrix}$$

$$Q_C = 1.6928 \quad \text{d.f.} = 6 \quad p = 0.95$$

$$\underset{\sim}{C}_2 = \begin{bmatrix} 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}$$

$$Q_C = 6.0962 \quad \text{d.f.} = 6 \quad p = 0.41$$

Neither of these hypotheses could be rejected ($\alpha = .05$), so it was concluded that the effect for time was not device-dependent, but in fact was uniform across the devices.

The generalized kappa statistic allows one to assess the extent of agreement for observers for a particular procedure. If one takes advantage of the modeling aspects of the GSK approach, one can also address the question of whether the same pattern of agreement holds from one time point to the next, and whether the same pattern holds for each of the procedures. The kappa statistic is of the form

$$\kappa = \frac{\pi_o - \pi_e}{1 - \pi_e}$$

where π_o is an observational probability of agreement and π_e is a hypothetical expected probability of agreement under an appropriate set of constraints --- here, the independence of observer classifications. The following illustrates the general form of the operations performed on a proportion vector in order to produce the kappa statistic:

$$\underset{\sim}{F}(\underset{\sim}{p}) = \exp[\underset{\sim}{A}_4(\log\{\underset{\sim}{A}_3[\exp(\underset{\sim}{A}_2\{\log \underset{\sim}{A}_1\underset{\sim}{p}\})]\})].$$

In order to generate kappa statistics for each procedure for each time, the proportion vector manipulated was that formed from the 7 × 16 contingency table whose seven rows of sixteen cells correspond to the seven 4 by 4 tables described above. The following linear operation matrices are those applied to each of the seven segments of the proportion vector calculated from the 7 × 16 table, resulting in kappa statistics, presented in Table 3, for Time 1 and Time 2 for each procedure.

$$\underset{\sim}{A}_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

$$\underset{\sim}{A}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & & & \\ 0 & 1 & 0 & 1 & 0 & & & \\ 0 & 1 & 0 & 0 & 1 & & \underset{\sim}{0} & \\ 0 & 0 & 1 & 1 & 0 & & & \\ 0 & 0 & 1 & 0 & 1 & & & \\ & & & & & 1 & 0 & 0 & 0 & 0 \\ & & & & & 0 & 1 & 0 & 1 & 0 \\ & & \underset{\sim}{0} & & & 0 & 1 & 0 & 0 & 1 \\ & & & & & 0 & 0 & 1 & 1 & 0 \\ & & & & & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

$$\underset{\sim}{A}_3 = \begin{bmatrix} 1 & -1 & 0 & 0 & -1 & & & \underset{\sim}{0} & \\ 0 & 0 & 1 & 1 & 0 & & & & \\ & & \underset{\sim}{0} & & & 1 & -1 & 0 & 0 & -1 \\ & & & & & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

$$\underset{\sim}{A}_4 = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

A weighted least squares analysis was applied to fit the model $\underset{\sim}{F}(\underset{\sim}{p}) \hat{=} \underset{\sim}{X}\underset{\sim}{b}$, where $F(\underset{\sim}{p})$ is the function vector consisting of the fourteen kappa statistics, and $\underset{\sim}{X}$ is the 14 × 14 identity matrix. The hypotheses tested for this model were:

1. There is no difference between the Time 1 and Time 2 kappa statistics across procedures.
2. There is no difference between the Time 1 kappa statistics among the procedures.
3. There is no difference between the Time 2 kappa statistics among the procedures.
4. The difference between the Time 1 kappa and the Time 2 kappa does not vary among procedures.

The appropriate $\underset{\sim}{C}$ matrix and chi-square statistics are reported in Table 4. The only hypothesis not rejected (α = .05 level of significance) was the fourth one, which had a chi-square of 5.81 (p value = .44). This suggests that a reduced model would be adequate. The 8-parameter model $\underset{\sim}{F}(\underset{\sim}{p}) \hat{=} \underset{\sim}{X}_2\underset{\sim}{b}$ was then fitted, where

$$\underset{\sim}{X}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

The goodness of fit was Q_C = 5.81 (df = 6), and p = .44, indicating an adequate fit. The parameter estimates and standard errors are summarized in Table 5. b_1 through b_7 can be interpreted as estimates of kappa for each procedure for Time 1, and b_8 is a common parameter reflecting the difference in the kappa statistic from Time 1 to Time 2 for all procedures. Thus, we can conclude that the kappa statistics are device-dependent and time-dependent. However, the change in the kappa statistic from Time 1 to Time 2 is homogeneous across devices.

EXAMPLE 2

The previous example was concerned with observer agreement where the primary unit of measurement was the averaged assessment of agreement for one device at one time point for one subject. Depending on whether the averaged assessment was positive or negative, the subject was classified into one of the cells of a four by four contingency table, whose entries were the basis of analysis. In this example, data are from Time 2 readings only, and the primary unit of measurement is the response profile of a pair of observers' readings for the same reaction; it will be --, -+, +-, or --. When the readings for both the standard and test reactions are considered, there are 4^2 = 16 possible response patterns. In this analysis, interest focuses on whether a pair of observers agree on a positive rating, agree on a negative rating, or disagree. The basis for analysis in this example is the two way classification of the pairs into such categories for both the standard and test procedure. The 3 × 3 table for the duplicate standard test procedure data is displayed below:

Test device: duplicate standard

Number of observers recording positive reaction on standard arm

Number of observers recording positive reaction on test arm

	⊖	×	⊕
⊖	675	29	23
×	39	14	14
⊕	27	19	250

Each circled '+' or '-' indicates that the observers agreed on that status, while the '×' indicates a disagreement. These classifications have been applied to the data for each procedure at Time 2.

Attention is first directed at assessing the extent of agreement between two observers on the two reactions. If d is the number of observers, there are d determinations for one reaction, and d on the other. If one assumes that each of two observers made the classifications independently of the other, the measuring process can be considered a series of independent Bernoulli trials. Let p_0 represent the probability of observers agreeing on a negative reading, p_1 be the probability of one positive reading, and p_2 be the probability of 2 positive readings. Let n_i, i = 0,1,2 be the number of pairs whose ratings include i positives - i.e. if both observers agree on a negative rating there are 0 positive ratings, if they disagree there is 1 positive rating, etc. The average probability of a positive reading for one pair of observers is

$$\bar{p} = 0\cdot\frac{n_0}{n} + \frac{n_1}{2n} + \frac{n_2}{n} = \frac{n_1}{2n} + \frac{n_2}{n} = \frac{1}{2}p_1 + p_2; \quad \bar{p} \text{ is}$$

used as the estimate of the true probability of a positive reading. Under independence, $E\{\hat{p}_0\} = (1 - \bar{p})^2$, and $E\{\hat{p}_2\} = \bar{p}^2$. Since $p_0 + p_1 + p_2 = 1$, $(1 - \bar{p}) = p_0 + \frac{1}{2}p_1$. Thus, the kappa statistic which applies here is

$$\kappa = \frac{p_0 + p_2 - \left(p_0 + \frac{p_1}{2}\right)^2 - \left(p_2 + \frac{p_1}{2}\right)^2}{1 - \left(p_0 + \frac{p_1}{2}\right)^2 - \left(p_2 + \frac{p_1}{2}\right)^2}$$

The 3 × 3 tables discussed above were analyzed over procedures by transforming each device's 3 × 3 table into an overall table with seven rows and nine columns. The 63 × 1 observed proportion vector was formed, and matrix

operations applied to generate these kappa statistics for each of the two tests of seven procedures at Time 2.

The form of the compounded function vector $\underset{\sim}{F}(\underset{\sim}{p})$ is as follows:

$$\exp\{\underset{\sim}{A}_4[\log[\underset{\sim}{A}_3\{\exp[\underset{\sim}{A}_2(\log(\underset{\sim}{A}_1\cdot\underset{\sim}{p}))]\}+\underset{\sim}{C}]]\}$$

The linear operator matrices for this transformation are:

$$\underset{\sim}{A}_1 = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & \frac{1}{2} & 0 & 1 & \frac{1}{2} & 0 & 1 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 1 & 0 & \frac{1}{2} & 1 & 0 & \frac{1}{2} & 1 \end{bmatrix} \otimes \underset{\sim}{I}_7$$

$$\underset{\sim}{A}_2 = \begin{bmatrix} 1 & & & & & \\ & 2 & & 0 & & \\ & & 2 & & & \\ & & & 1 & & \\ & 0 & & & 2 & \\ & & & & & 2 \end{bmatrix} \otimes \underset{\sim}{I}_7$$

$$\underset{\sim}{A}_3 = \begin{bmatrix} 1 & -1 & -1 & & & 0 \\ 0 & -1 & -1 & & & \\ & 0 & & 1 & -1 & -1 \\ & & & 0 & -1 & -1 \end{bmatrix} \otimes \underset{\sim}{I}_7$$

$$\underset{\sim}{A}_4 = [1 \; -1] \otimes \underset{\sim}{I}_7$$

The constant vector $\underset{\sim}{C}$ added was $\underset{\sim}{C} = [0\ 1] \otimes \underset{\sim}{I}_{14}$.

The fourteen kappa statistics and their standard errors are found in Table 6. The model $\underset{\sim}{F}(\underset{\sim}{p}) \hat{=} \underset{\sim}{X}\underset{\sim}{b}$ was fitted with the 14 × 14 identity matrix as the design matrix. Two hypotheses were tested:

1. There is no difference in the kappa statistics for the standard procedures.
2. There is no difference in the kappa statistics for the test procedures.

Hypothesis 1 was not rejected (p = .94) and Hypothesis 2 was rejected (p = .00). The corresponding contrast matrices and chi-square statistics are also found in Table 6. A 2-parameter, reduced model was then fitted. The design matrix is as follows:

$$\underset{\sim}{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$

The resulting goodness of fit statistic was 4.2841 (12 degrees of freedom), with p-value of 0.98, indicating a good fit. The estimate of b_1 is 0.8620 ± 0.0061 and the estimate of b_2 is -0.0899 ± 0.0100. The parameter b_1 can be interpreted as being the kappa statistic for the standard procedure, and b_2 represents the difference between the statistic for the standard and test procedure for each of the seven groups except for the first one. However, that test procedure is the duplicate standard, which, understandably, has an estimate nearly identical to that of the standard. It is concluded therefore that the extent of agreement between two observers was similar for the standard procedures, and also similar for the test procedures as well, there being a constant difference between the standard kappa statistic and the test kappa statistic, except for the case discussed above.

The analysis of the seven by nine table is continued by examining summary measures of the extent of agreement in the standard test and the experimental test for the seven procedures. In order to investigate the pattern of reliability in the data, a short series of hierarchical kappa statistics was computed. A hierarchy of weights is used to combine certain adjacent response categories in order to create successively less stringent definitions of agreement. These weighted kappa statistics provide a framework for investigating the internal mechanisms that contribute to the decreasing reliability resulting from the broader definitions of judgment criteria.

The two sets of weights used in the analysis are displayed in Table 7. If $n_{jj'}$ denotes the number of subjects in the j-th response category for the standard procedure and the j'th category for the test procedure (for j, j' = 0,1,2 for no positives, one positive, or two positives), then the weighted kappa statistic created can be written as

$$\kappa = \frac{\pi_0 - \pi_E}{1 - \pi_E} \text{ where } \pi_0 \hat{=} \sum_{j=0}^{2}\sum_{j'=0}^{2} w_{h,jj'}(n_{jj'}/n)$$

is the observed agreement and

$\pi_E \hat{=} \sum_{j=0}^{2}\sum_{j'=0}^{2} w_{h,jj'}(n_{j+}n_{+j'}/n^2)$ is the agreement expected under independence.

The analysis begins with the 63 × 1 observed proportion vector.

The compounded function vector $\underset{\sim}{F}(\underset{\sim}{p})$ has the same general form as it had in Example 1 (p. 1). The necessary linear operator matrices are:

$$\underset{63 \times 63}{A_1} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \end{bmatrix} \otimes \underset{\sim}{I}_7$$

$$\underset{77 \times 56}{\underset{\sim}{A}_2} = \begin{bmatrix} 1&0&0&1&0&0&0&0 \\ 1&0&0&0&1&0&0&0 \\ 1&0&0&0&0&1&0&0 \\ 0&1&0&1&0&0&0&0 \\ 0&1&0&0&1&0&0&0 \\ 0&1&0&0&0&1&0&0 \\ 0&0&1&1&0&0&0&0 \\ 0&0&1&0&1&0&0&0 \\ 0&0&1&0&0&1&0&0 \\ 0&0&0&0&0&0&1&0 \\ 0&0&0&0&0&0&0&1 \end{bmatrix} \otimes \underset{\sim}{I}_7$$

$$\underset{28 \times 77}{\underset{\sim}{A}_3} = \begin{bmatrix} -1&0&0&0&-1&0&0&0&-1&1&0 \\ -1&0&0&0&-1&-1&0&-1&-1&0&1 \\ 0&1&1&1&0&1&1&1&0&0&0 \\ 0&1&1&1&0&0&1&0&0&0&0 \end{bmatrix} \otimes \underset{\sim}{I}_7$$

$$\underset{14 \times 28}{\underset{\sim}{A}_4} = \begin{bmatrix} 1&0&-1&0 \\ 0&1&0&-1 \end{bmatrix} \otimes \underset{\sim}{I}_7$$

The estimates and standard errors for these kappa statistics are shown in Table 8. These values reflect the expected increase in agreement for the second set of weights. Weighted least squares analyses would also be appropriate here as a way to explore the pattern of variation in these measures of agreement in a manner similar to those described above.

REFERENCES

Grizzle, J. E., Starmer, C. F. and Koch, G. G. (1969). Analysis of categorical data by linear models. Biometrics 25, 489-504.

Koch, G. G., Landis, J. R., Freeman, J. L., et al. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. Biometrics 33, 133-158.

Koch, G. G. (1981). Hierarchical kappa statistics. To appear in Encyclopedia of Statistical Sciences, Norman L. Johnson and Samuel Kotz, eds., to be published by John Wiley and Sons, Inc.

Landis, J. R. and Koch, G. G. (1975). A review of statistical methods in the analysis of data arising from observer reliability studies, Parts I and II. Statistica Neerlandica 29, 101-123, 151-161.

Landis, J. R., Stanish, W. M., Freeman, J. L. and Koch, G. G. (1976). A computer program for the generalized chi-square analysis of categorical data using weighted least squares (GENCAT). Computer Programs in Biomedicine 6, 196-231.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics 33, 159-174.

Landis, J. R. and Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 33, 363-374.

ACKNOWLEDGEMENTS

This research was supported in part by the U.S. Bureau of the Census (JSA-80-19). The authors would like to express their appreciation to Ann Thomas and Lori Turnbull for their typing of the manuscript.

TABLE 1: FREQUENCIES BY STANDARD READING AND TEST READING FOR TIMES 1 AND 2, FOR SUBJECTS RECEIVING DUPLICATE STANDARD AS THE TEST PROCEDURE (EXAMPLE 1)

			Reading at Time 2				
	std		-	-	+	+	Total
		dup	-	+	-	+	
Reading at Time 1	-	-	509	4	17	3	533
	-	+	13	8	0	8	29
	+	-	14	1	17	9	41
	+	+	7	4	9	170	190
	Total		543	17	43	190	793

TABLE 2: PARAMETERS AND STANDARD ERRORS FOR TIME CHANGE FOR CONTRA-NEGATIVE STANDARD AND FOR CONTRA-POSITIVE STANDARD (EXAMPLE 1)

Device	Time change for contra-negative standard (and standard errors)		Time change for contra-positive standard (and standard errors)	
A	0.021	(0.010)	-0.007	(0.027)
B	0.026	(0.016)	0.055	(0.027)
C	0.031	(0.013)	0.059	(0.033)
D	0.033	(0.016)	0.012	(0.024)
E	0.011	(0.014)	-0.021	(0.031)
F	0.032	(0.017)	0.013	(0.026)
G	0.025	(0.015)	0.026	(0.025)

TABLE 3: KAPPA STATISTICS (AND STANDARD ERRORS) FOR EACH DEVICE AND EACH TIME POINT (EXAMPLE 1)

Device	Time	Kappa	s.e.
A	1	0.783	0.025
	2	0.812	0.023
B	1	0.568	0.032
	2	0.643	0.030
C	1	0.510	0.037
	2	0.607	0.035
D	1	0.566	0.030
	2	0.612	0.029
E	1	0.644	0.030
	2	0.642	0.030
F	1	0.525	0.029
	2	0.573	0.028
G	1	0.611	0.030
	2	0.668	0.028

TABLE 5: PARAMETER ESTIMATES AND STANDARD ERRORS FOR REDUCED MODEL FOR KAPPA STATISTICS (EXAMPLE 1)

$$\underset{\sim}{b} = \begin{bmatrix} 0.772 \pm 0.021 \\ 0.620 \pm 0.026 \\ 0.565 \pm 0.026 \\ 0.525 \pm 0.026 \\ 0.537 \pm 0.032 \\ 0.616 \pm 0.026 \\ 0.583 \pm 0.028 \\ 0.048 \pm 0.011 \end{bmatrix}$$

TABLE 4: CONTRAST MATRICES AND CHI-SQUARE STATISTICS FOR TESTS OF HYPOTHESES CONCERNING KAPPA STATISTICS (EXAMPLE 1)

1. H_0: no difference between the Time 1 and Time 2 kappa statistics across procedures.

$$\underset{\sim}{C}_1 = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

$Q_C = 24.37$ d.f. = 7 p = 0.0

2. H_0: no difference between the Time 1 kappa statistics among the procedures.

$$\underset{\sim}{C}_2 = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \end{bmatrix}$$

$Q_C = 70.117$ d.f. = 6 p = 0

3. H_0: no difference between the Time 2 kappa statistics among the procedures.

$$\underset{\sim}{C}_3 = \begin{bmatrix} 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \end{bmatrix}$$

$Q_C = 57.544$ d.f. = 6 p = 0

4. H_0: no variation across procedures in the difference between Time 1 kappa and Time 2 kappa.

$$\underset{\sim}{C}_4 = \begin{bmatrix} 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

$Q_C = 5.81$ d.f. = 6 p = .44

TABLE 6A: KAPPA STATISTICS AND STANDARD ERRORS BASED ON INDIVIDUAL OBSERVER ASSESSMENTS (EXAMPLE 2)

Device Used as Test	Procedure	Kappa	s.e.
A	std.	0.854	0.017
	test	0.865	0.017
B	std.	0.862	0.017
	test	0.759	0.021
C	std.	0.858	0.018
	test	0.790	0.022
D	std.	0.861	0.017
	test	0.759	0.021
E	std.	0.850	0.017
	test	0.786	0.021
F	std.	0.878	0.016
	test	0.761	0.019
G	std.	0.866	0.017
	test	0.779	0.020

TABLE 6B: HYPOTHESES, CONTRAST MATRICES, AND TESTS OF SIGNIFICANCE FOR KAPPA STATISTICS SHOWN IN 7A* (EXAMPLE 2)

H_0: no difference in standard device agreements

$$\underset{\sim}{C}_1 = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \end{bmatrix}$$

$Q_C = 1.771$ d.f. = 6 p = 0.94

H_0: no difference in test device agreements

$$\underset{\sim}{C}_2 = \begin{bmatrix} 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \end{bmatrix}$$

$Q_C = 26.647$ d.f. = 6 p = 0.002

*Design matrix is 14 × 14 identity matrix.

TABLE 7: CRITERION WEIGHTS FOR HIERARCHICAL KAPPA STATISTICS (EXAMPLE 2)

Criterion (h)	1	2	3	4	5	6	7	8	9
1	1	0	0	0	1	0	0	0	1
2	1	0	0	0	1	1	0	1	1

TABLE 8: ESTIMATES AND STANDARD ERRORS FOR UNWEIGHTED AND WEIGHTED KAPPA STATISTICS (EXAMPLE 2)

Device	Criterion Weight	Kappa	s.e.
A	1	0.706	0.021
	2	0.754	0.021
B	1	0.525	0.024
	2	0.549	0.027
C	1	0.480	0.026
	2	0.500	0.029
D	1	0.535	0.023
	2	0.549	0.026
E	1	0.604	0.023
	2	0.642	0.025
F	1	0.493	0.022
	2	0.512	0.024
G	1	0.552	0.023
	2	0.587	0.025