Measurement System Analysis for Binary Data

Wessel N. van Wieringen

Department of Mathematics
Vrije Universiteit Amsterdam
1081 HV Amsterdam, The Netherlands
(*wvanwie@few.vu.nl*)

Jeroen de Mast

Institute for Business and Industrial Statistics
of the University of Amsterdam (IBIS UvA)
1018 TV Amsterdam, The Netherlands
(*j.demast@uva.nl*)

We describe a methodology for the assessment of the repeatability and reproducibility (R&R) of measurement systems that measure on a binary scale, such as pass–fail inspections. We focus on the situation where no reference values can be obtained for the objects in the experiment and consequently model the results of the R&R experiment as a latent class model. We provide estimators based on the maximum likelihood approach and the method of moments, and compare their properties. We also give guidelines for model checking and recommendations for sample sizes. The methodology is illustrated by an example.

KEY WORDS: Categorical data; Gauge capability; Latent class model; Reliability; Repeatability and reproducibility.

1. INTRODUCTION

Measurement system analysis (MSA) describes, categorizes, and evaluates the quality of measurements (cf. Allen and Yen 1979). An important aspect of the quality of measurements is their repeatability and reproducibility (R&R), referring to the extent to which similar results are obtained when the same object is measured multiple times. The degree of similarity of measurements done by the same rater and under identical circumstances is the repeatability, whereas reproducibility refers to the degree of similarity when multiple raters perform the measurements, possibly under varying conditions. R&R is assessed from an experiment, designed to quantify the effect of factors onto the variability of the measurements. In the standard-gauge R&R experiment (Burdick, Borror, and Montgomery 2003), applicable to measurements on a numerical scale, the experimental setup involves the factors objects and raters.

In this article we study the assessment of the R&R of measurements on a binary scale that have only two values, such as *pass* and *fail*. The omnipresence of pass/fail and other binary inspection systems in industry makes a sound methodology to study their R&R an important subject of study. We focus on the situation in which a so-called "gold standard" is not available, that is, the true state of the objects included in the experiment (e.g., "defective" or "functional") cannot be known. Boyles (2001) described the evaluation of binary measurement systems. In his experimental setup, a single measurement system (i.e., one rater) rates each object multiple times (enabling the assessment of repeatability only). Boyles used a latent class model to model the outcome of the experiment. Latent class models describe experiments with categorical outcomes and assume an underlying, unobservable (latent) categorical variable, which is used to explain the structure in the observed data (Bartholomew and Knott 1999). Boyles (2001) estimated the parameters of the latent class model by the maximum likelihood method. The same model and estimation method was described briefly by Van Wieringen and Van den Heuvel (2005), who compared them to alternative approaches.

The standard experimental design considered in the biostatistical literature (e.g., Hui and Walter 1980; Garrett, Eaton, and Zeger 2002) involves two factors, namely objects and tests (similar to our factor raters). Repetitions are not considered in the biostatistical literature, however. Latent class models are the basis of analysis in the biostatistical literature as well. An important application of R&R studies for binary measurements in biostatistics is in evaluating the performance of diagnostic tests in the absence of a gold standard.

In this article we extend the methods proposed by Hui and Walter (1980), Boyles (2001), Van Wieringen and Van den Heuvel (2005), and others, in the sense that our setup involves both factors objects and raters, as well as repetitions (multiple raters judge each object multiple times), thus allowing the assessment of both repeatability and reproducibility. The latent class models provided in the literature are modified to describe such experiments. In industry, the possibility to assess both R&R components of measurement spread is considered important, because it provides vital clues needed to improve the reliability of the measurements. We describe the maximum likelihood method for estimating the parameters. We also outline estimation by the method of moments (not given in Boyles 2001 or Van Wieringen and Van den Heuvel 2005). Normal approximations to the (asymptotic) variance of both estimators are given, and the two estimation methods are compared in a simulation study. Assessment of the validity of the models, an issue not covered by Boyles (2001) and Van Wieringen and Van den Heuvel (2005), is a crucial aspect of latent class modeling. We propose diagnostics for the latent class model. As a final contribution, we offer sample size recommendations (obtained from additional simulations) and an example illustrating the proposed method.

Thus the purpose of this article is to generalize methods proposed earlier to be able to cope with the important situation in which both repeatability and reproducibility are to be estimated, and to consummate these methods by outlining how to obtain diagnostics and standard errors for the estimates and by supplying guidelines for sample sizes.

2. THE MODEL

The experimental design for evaluating the R&R of a binary measurement system involves n objects that are judged repeatedly by m raters. The data from the experiment are denoted by

TECHNOMETRICS, NOVEMBER 2008, VOL. 50, NO. 4
DOI 10.1198/004017008000000415

X_{ijh}, with $i = 1, \ldots, n$ and $j = 1, \ldots, m$ indexing the objects and raters. We denote repeated judgments by $h = 1, \ldots, \ell$; note that the estimation methods discussed later can be generalized to situations where the number of repetitions is not the same for all raters. The X_{ijh}'s are 0 ("reject") or 1 ("pass"). We assume that the "true" (henceforth called "reference") value of the measured object, Y_i, is also either 0 ("defect") or 1 ("good"), and we write $\theta = P(Y_i = 1)$ for the proportion of good objects. Further, we assume that, conditional on Y_i, the $\{X_{ijh}\}_{j,h}$'s are independent. This is the assumption of conditional independence, standard in latent class models, which can be formulated as

$$P(X_{i,1,1}, X_{i,1,2}, \ldots, X_{i,m,\ell}|Y_i) = \prod_{j,h} P(X_{ijh}|Y_i).$$

The measurement X_{ijh} depends on Y_i, and we define $\pi_j(y) = P(X_{ijh} = 1|Y_i = y)$. We have the unconditional distribution

$$\begin{aligned} P(X_{ijh} = x) &= P(X_{ijh} = x|Y_i = 0)P(Y_i = 0) \\ &\quad + P(X_{ijh} = x|Y_i = 1)P(Y_i = 1) \\ &= (1-\theta)\pi_j(0)^x(1-\pi_j(0))^{1-x} \\ &\quad + \theta\pi_j(1)^x(1-\pi_j(1))^{1-x}, \end{aligned}$$

where $x = 0, 1$. The latent class model distinguishes between a manifest variable (the measurement of a rater) and an unobserved, latent variable (the reference value of the object, which is assumed to be unknown). The latter is used to explain the correlation structure in the observations (namely, that ratings of the same object are correlated, unless the ratings are done completely at random). In line with Hui and Walter (1980) and Boyles (2001), this approach is built around the notion that the latent variable is binary as well, and, in combination with the conditional independence assumption, it assumes that the population of objects divides into two subpopulations (of good and defective objects), and that these subpopulations are homogeneous, in the sense that $P(X_{ijh} = x)$ depends only on Y_i and not on other characteristics of the object. The assumption may be violated in cases where the underlying characteristic is a continuum (such as a continuous quality characteristic), rather than a dichotomy (good or defective). In this case $P(X_{ijh} = x)$ is not equal within the subpopulations of good and defective objects, and the conditional independence assumption is violated. We recommend an approach based on a latent trait (instead of latent class) model in that case, such as that described by De Mast and Van Wieringen (2008). Note that in a later section we describe diagnostic checks that allow the user to verify whether the conditional independence assumption holds; we demonstrate the effectiveness of these diagnostic checks from an example. In addition, by carefully sampling the objects for the MSA experiment, ensuring conditional exchangeability, the user achieves robustness against violations of conditional independence (as described later).

We let $\mathbf{X}$ we denote the matrix containing the experimental outcomes, which we rewrite in terms of response patterns. Let $\mathbf{R} = (R_{ij})_{i=1,\ldots,n;j=1,\ldots,m}$, with $R_{ij} = \sum_{h=1}^{\ell} X_{ijh}$. The likelihood function is

$$\begin{aligned} L(\mathbf{R}; \boldsymbol{\Psi}) = \prod_{i=1}^{n} \Bigg(& (1-\theta) \prod_{j=1}^{m} \binom{\ell}{R_{ij}} (1-\pi_j(0))^{\ell - R_{ij}} (\pi_j(0))^{R_{ij}} \\ & + \theta \prod_{j=1}^{m} \binom{\ell}{R_{ij}} (1-\pi_j(1))^{\ell - R_{ij}} (\pi_j(1))^{R_{ij}} \Bigg), \end{aligned} \quad (1)$$

where $\boldsymbol{\Psi} = (\theta, \pi_1(1), \ldots, \pi_m(1), \pi_1(0), \ldots, \pi_m(0))^T$.

To ensure identifiability, it is sufficient to require $\theta \in (0, 1)$, $1 \geq \pi_j(1) > \pi_j(0) \geq 0$ for all j, and

$$(\ell + 1)^m - 1 \geq 2m + 1. \quad (2)$$

(A more general form was proved in Van Wieringen 2005.)

This model treats the differences among raters as fixed effects, because the parameters $\pi_j(1)$ and $\pi_j(0)$ reflect characteristics of the raters individually, not of a population of raters. Often there is a single rater (who could be a person, but also an automatic device) for each of a limited number of production lines; here one would typically include all raters in the experiment, and a model with fixed rater effects is appropriate. On the other hand, if the population of raters is large, then one takes a sample of the raters and fits a random-effects model, as is done by Qu, Tan, and Kutner (1996) for the situation where $\ell = 1$. But even in this case, where the raters are a sample from a larger population, there is an argument that can be made for a fixed-effects model. The number of raters that should be included in the experiment to allow a reliable estimation of the random-effects model is so large as to make it practically impossible in general. One work-around is to include far less raters than needed, and accept that confidence intervals are very wide. This strategy is common practice in industry (see Vardeman and Van Valkenburg 1999 for numerical R&R studies). An alternative strategy is to be more modest about the generalizability of one's conclusions and accept that the sample of raters typically is too small to allow reliable inferences about the population of raters. In that case, one fits a fixed-effects model and focuses on solving the inspection problems encountered by the raters in the experiment, hoping that improvement actions will have a positive spinoff for the other raters as well.

The latent class model allows a natural operationalization of the R&R of the measurements. For binary ratings, measurement error comes down to misclassification, and we propose to express R&R as a probability of misclassification. From biostatistics, we adopt the terms "sensitivity" and "specificity." The sensitivity of rater j is defined as $\pi_j(1)$, the probability that a good object passes. Rater j's specificity is $1 - \pi_j(0)$, the probability that a defective object is failed. Sensitivity and specificity are the complements of type I error and type II error. Given the distribution θ of the objects and estimates $\hat{\pi}_1(0), \ldots, \hat{\pi}_m(0)$ and $\hat{\pi}_1(1), \ldots, \hat{\pi}_m(1)$, and assuming that all raters measure the same number of objects, the estimated probability of misclassification is

$$P(\text{misclassification}) = \frac{1}{m} \sum_{j=1}^{m} \big(\theta(1 - \hat{\pi}_j(1)) + (1-\theta)\hat{\pi}_j(0)\big). \quad (3)$$

If raters measure different numbers of objects, then minor modifications are required. In addition, for a particular object, one can indicate from which category the object is most likely to originate, namely the category y that maximizes $P(Y_i = y|R_{i1}, R_{i2}, \ldots, R_{im})$.

3. ESTIMATION

In this section we describe how the parameters of the latent class model can be estimated using the maximum likelihood (ML) method and the method of moments (MoM).

3.1 Maximum Likelihood Method

The EM algorithm approaches the problem of maximizing the likelihood function indirectly by exploiting the more convenient form of a related likelihood. The related likelihood is obtained by rewriting the likelihood assuming that we know the actual values of the reference values Y_i, which is called the "complete" likelihood function. It can be shown that the ML estimator produced by the EM algorithm maximizes not only the complete likelihood function L_c, but also the original likelihood function L related to the "incomplete" data (i.e., where the Y_i's are unknown). McLachlan and Krishnan (1997) have given a general account of ML estimation with the EM algorithm with many properties, generalizations, and applications.

The complete likelihood is given by

$$L_c((\mathbf{R}, \mathbf{Y}); \boldsymbol{\Psi}) = \prod_{i=1}^{n} \{(1-\theta) P_{\boldsymbol{\Psi}}(\mathbf{R}_i = \mathbf{r}_i|Y_i)\}^{1-Y_i} \times \{\theta P_{\boldsymbol{\Psi}}(\mathbf{R}_i = \mathbf{r}_i|Y_i)\}^{Y_i}.$$

Taking the logarithm, we get

$$\begin{aligned}\log\big(L_c((\mathbf{R}, \mathbf{Y}); \boldsymbol{\Psi})\big) &\propto \sum_{i=1}^{n} \Bigg((1-Y_i)\log(1-\theta) + Y_i \log(\theta) + (1-Y_i) \\ &\quad \times \sum_{j=1}^{m} \big(R_{ij}\log(\pi_j(0)) + (\ell - R_{ij})\log(1-\pi_j(0))\big) \\ &\quad + Y_i \sum_{j=1}^{m} \big(R_{ij}\log(\pi_j(1)) + (\ell - R_{ij})\log(1-\pi_j(1))\big) \Bigg), \quad (4)\end{aligned}$$

a convenient form when it comes to maximization.

The EM algorithm, applied to the estimation of $\boldsymbol{\Psi}$ in model (1), can be described as follows (McLachlan and Krishnan 1997):

Step 1. Choose initial values for the estimate $\hat{\boldsymbol{\Psi}}^{(0)}$, and specify a stopping criterion.

Step 2 (E-step). We have no knowledge of $\mathbf{Y}$; however, using the current estimate of $\hat{\boldsymbol{\Psi}}^{(t)}$, we replace $\mathbf{Y}$ by its conditional expectation given $\mathbf{R}$,

$$\hat{Y}_i = E_{\hat{\boldsymbol{\Psi}}^{(t)}}(Y_i|\mathbf{R}).$$

In fact, real values from the interval [0, 1] are substituted for $\mathbf{Y}$, whereas their proper value is either 0 or 1. Furthermore, using Bayes's theorem,

$$\begin{aligned}E_{\hat{\boldsymbol{\Psi}}^{(t)}}(Y_i|\mathbf{R}) &= \sum_{y=0}^{1} y P_{\hat{\boldsymbol{\Psi}}^{(t)}}(Y_i = y|\mathbf{R}) = P_{\hat{\boldsymbol{\Psi}}^{(t)}}(Y_i = 1|\mathbf{R}) \\ &= \big(P_{\hat{\boldsymbol{\Psi}}^{(t)}}(\mathbf{R}|Y_i = 1) P_{\hat{\boldsymbol{\Psi}}^{(t)}}(Y_i = 1)\big) / \big(P_{\hat{\boldsymbol{\Psi}}^{(t)}}(\mathbf{R})\big).\end{aligned}$$

Calculating this expectation we find that

$$\begin{aligned}\hat{Y}_i^{(t)} &= \hat{\theta}^{(t)} P_{\hat{\boldsymbol{\Psi}}^{(t)}}(\mathbf{R}_i = \mathbf{r}_i|Y_i = 1) \\ &\quad / \big(\hat{\theta}^{(t)} P_{\hat{\boldsymbol{\Psi}}^{(t)}}(\mathbf{R}_i = \mathbf{r}_i|Y_i = 1) \\ &\quad + \big(1 - \hat{\theta}^{(t)}\big) P_{\hat{\boldsymbol{\Psi}}^{(t)}}(\mathbf{R}_i = \mathbf{x}_i|Y_i = 0)\big). \quad (5)\end{aligned}$$

Step 3 (M-step). The M-step consists of maximizing the complete log-likelihood function (4). Taking the first-order partial derivatives, equating them to 0, and solving these equations with respect to the parameters yields the estimates

$$\hat{\theta}^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \hat{Y}_i^{(t)},$$

$$\hat{\pi}_j^{(t+1)}(1) = \frac{\sum_{i=1}^{n} R_{ij} \hat{Y}_i^{(t)}}{\ell \sum_{i=1}^{n} \hat{Y}_i^{(t)}},$$

and

$$\hat{\pi}_j^{(t+1)}(0) = \frac{\sum_{i=1}^{n} R_{ij} (1 - \hat{Y}_i^{(t)})}{\ell \sum_{i=1}^{n} (1 - \hat{Y}_i^{(t)})}.$$

Thus the next estimate $\hat{\boldsymbol{\Psi}}^{(t+1)}$ of the parameters $\boldsymbol{\Psi}$ is obtained.

Step 4. Go back to step 2 until the stopping criterion has been satisfied.

We establish the variance of the ML estimator by means of the normal approximation. To obtain the Fisher information matrix, $I(\mathbf{R}; \boldsymbol{\Psi})$, of the incomplete information, we use the following relation, given by McLachlan and Krishnan (1997):

$$\begin{aligned}I(\mathbf{R}; \boldsymbol{\Psi}) &= -\frac{\partial^2}{\partial \boldsymbol{\Psi}\, \partial \boldsymbol{\Psi}^T} \log(L(\mathbf{R}; \boldsymbol{\Psi})) \\ &= -E_{\boldsymbol{\Psi}}\left(\frac{\partial^2}{\partial \boldsymbol{\Psi}\, \partial \boldsymbol{\Psi}^T} \log\left(\frac{L_c((\mathbf{R}, \mathbf{Y}); \boldsymbol{\Psi})}{L(\mathbf{R}; \boldsymbol{\Psi})} \right) \Big| \mathbf{R} = \mathbf{r} \right) \\ &\quad - \operatorname{cov}_{\boldsymbol{\Psi}}\big(\nabla_{\boldsymbol{\Psi}} \log\big(L_c((\mathbf{R}, \mathbf{Y}); \boldsymbol{\Psi})\big) | \mathbf{R} = \mathbf{r} \big).\end{aligned}$$

This relation allows a straightforward calculation (Van Wieringen 2003) of the observed Fisher information matrix and, consequently, the construction of confidence regions for the estimates.

The confidence intervals also may be constructed empirically, following, for instance, De Menezes (1999), using the bootstrap. New experimental data are generated by randomly drawing (with replacement) n samples (objects) from the original experimental data. The parameters are estimated for the new data. The process (resampling and estimation) is repeated numerous, say B, times. The limits of the 95% confidence interval of the parameter estimates are given by the .025 and .975 quantiles of each set of B estimated parameters. Alternatively, the

profile likelihood approach suggested by Boyles (2001) may be used.

These approaches may not result in reliable confidence intervals of the parameter estimates in small sample size situations with near-perfect sensitivity and specificity. We have found that in the case where $\theta = .5$, $\pi_j(1) = .975$, and $\pi_j(0) = .025$, at least $n = 100$ (and preferably more) is needed for these procedures to yield reliable confidence interval estimates, based on an experiment with three raters and assuming (in line with our recommendation made later) $\ell = 3$ repetitions per rater.

3.2 Method of Moments

The MoM requires that the model parameters be expressed in terms of the moments of the distribution, whereupon estimators for the moments can be substituted in these expressions. In constructing moment estimators, we need the concept of mixed factorial moments. If $R_{i1}, \ldots, R_{im}$ are random variables, then we define their mixed factorial moments, for $a_1, \ldots, a_m = 0, 1, 2, \ldots,$ as

$$\mu_{(a_1,\ldots,a_m)} = E\left(\prod_{j=1}^{m} \frac{R_{ij}(R_{ij}-1)\cdots(R_{ij}-a_j+1)}{\ell(\ell-1)\cdots(\ell-a_j+1)}\right). \qquad (6)$$

It can be shown (Van Wieringen 2005) that the mixed factorial moments of $R_{i1}, \ldots, R_{im}$, distributed as in (1), are related to the parameters as $\mu_{(a_1,\ldots,a_m)} = (1-\theta)\prod_{j=1}^{m} \pi_j^{a_j}(0) + \theta \prod_{j=1}^{m} \pi_j^{a_j}(1)$ if $0 \le a_j \le \ell$ for all j, and $\mu_{(a_1,\ldots,a_m)} = 0$ if there is a j such that $a_j > \ell$.

The mixed factorial moment $\mu_{(a_1,\ldots,a_m)}$ is estimated by the mixed factorial sample moment, defined by

$$\hat{\mu}_{(a_1,\ldots,a_m)} = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{m} \frac{R_{ij}(R_{ij}-1)\cdots(R_{ij}-a_j+1)}{\ell(\ell-1)\cdots(\ell-a_j+1)}.$$

Henceforth, factorial moments are denoted with the use of unit vectors $\mathbf{e}_j = (0, \ldots, 0, 1, 0, \ldots, 0)$ (with all entries 0, except for the jth, which is 1), for example

$$\mu_{2\mathbf{e}_1+\mathbf{e}_2} = \mu_{(2,1,0,\ldots,0)} = E\left(\frac{R_{i1}(R_{i1}-1)R_{i2}}{\ell(\ell-1)\ell}\right).$$

When dealing with one rater, the identifiability restrictions (2) dictate that the rater measures at least three times ($\ell \ge 3$). Assuming that this holds, we express the parameters in terms of mixed factorial moments. For this particular situation, this has been done by Blischke (1962). Because this case is illustrative for other numbers of raters, we repeat it here. After algebraic manipulation of the mixed factorial moments, and exploiting their relationship with the parameters, we arrive at

$$\begin{aligned}\pi_1(1) + \pi_1(0) &= \left(\mu_{3\mathbf{e}_1} - \mu_{2\mathbf{e}_1}\mu_{\mathbf{e}_1}\right)/\left(\mu_{2\mathbf{e}_1} - \mu_{\mathbf{e}_1}\mu_{\mathbf{e}_1}\right)\\ &= b_1,\\ \pi_1(1)\pi_1(0) &= \left(\mu_{3\mathbf{e}_1}\mu_{\mathbf{e}_1} - \mu_{2\mathbf{e}_1}\mu_{2\mathbf{e}_1}\right)/\left(\mu_{2\mathbf{e}_1} - \mu_{\mathbf{e}_1}\mu_{\mathbf{e}_1}\right)\\ &= b_2,\end{aligned}$$

and

$$\theta = \left(\mu_{\mathbf{e}_1} - \pi_1(0)\right)/(\pi_1(1) - \pi_1(0)).$$

From these equations, we can solve, for θ, $\pi_1(1)$, and $\pi_1(0)$,

$$\pi_1(1) = \frac{1}{2}\left(b_1 \pm \sqrt{b_1^2 - 4b_2}\right),$$

$$\pi_1(0) = \frac{1}{2}\left(b_1 \mp \sqrt{b_1^2 - 4b_2}\right),$$

and

$$\theta = \left(2\mu_{\mathbf{e}_1} - b_1 \pm \sqrt{b_1^2 - 4b_2}\right)\Big/\left(\pm 2\sqrt{b_1^2 - 4b_2}\right).$$

Taking $\pi_1(1) > \pi_1(0)$, in line with the identifiability restrictions, only one solution for each parameter remains, fixing the solution for θ. Resubstituting the explicit form of the mixed factorial moments, given just after equation (6), proves the identity. To obtain estimates for the parameters, we replace the mixed factorial moments by the corresponding mixed factorial sample moments.

Analogous to the one-rater case, we can express the parameters of the more-than-one-rater latent class model in terms of mixed factorial moments. Details have been given by Van Wieringen (2005).

The variance of the MoM estimator is established by means of the normal approximation. Hereafter, we let $\hat{\boldsymbol{\Psi}}(\hat{\boldsymbol{\mu}})$ denote the estimate of the parameters $\boldsymbol{\Psi} = (\Psi_1, \ldots, \Psi_{2m+1})$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{2m+1})$ is the vector containing the $2m + 1$ mixed factorial moments used in constructing the estimators for the $2m + 1$ parameters and $\hat{\boldsymbol{\mu}}$ the estimate of $\boldsymbol{\mu}$. It can be shown (Van Wieringen 2003) that the variance of these moment estimators is asymptotically normally distributed as $\mathrm{N}(\boldsymbol{\Psi}(\boldsymbol{\mu}), n^{-1}\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T)$, where, in line with Serfling (1980),

$$\mathbf{D} = \left[\left.\frac{\partial \Psi_s}{\partial \hat{\mu}_t}\right|_{\hat{\boldsymbol{\mu}}=\boldsymbol{\mu}}\right]_{\substack{1\le s\le 2m+1\\ 1\le t\le 2m+1}}$$

and

$$\boldsymbol{\Sigma} = [\mathrm{cov}(\hat{\mu}_s, \hat{\mu}_t)]_{\substack{1\le s\le 2m+1\\ 1\le t\le 2m+1}}.$$

Explicit expressions of the terms in the covariance matrix have been given by Van Wieringen (2003). This normal approximation can be used in constructing confidence regions for the estimates. Alternatively, as for the ML method, confidence intervals may be constructed empirically, using a resampling technique like the bootstrap, as was done by De Menezes (1999).

4. COMPARISON OF THE ESTIMATION METHODS

To study the performance of the two estimation methods described earlier, we conducted a simulation experiment for the situations involving one, two, three, and four raters (situations in which the rater effect often is considered fixed). We varied the number of objects n from 20 to 150, the number of repetitions ℓ from 2 (if identifiability restrictions allowed) to 15, and the parameters θ from .50 to .90, $\pi_j(1)$ from .55 to .95, and $\pi_j(0)$ from .10 to .40. The simulation consists of generating realizations, $\mathbf{X}$, from the distribution (1), followed by estimating the parameters by ML and MoM. For any specific choice of n, ℓ, and $\boldsymbol{\Psi}$, we generated 10,000 realizations, each resulting in estimates for the parameters. From these 10,000 estimates, we computed the average and standard deviation for

each parameter. We carried out the simulations with Matlab version 6.5.0.189013a, release 13. What can be observed in the simulation results can be summarized as follows:

- The bias and standard deviation of the estimates of both methods are comparable when $\theta \approx .5$ and $\pi_j(1) \gg \pi_j(0)$. This is the most relevant situation. Any practical inspection system will operate with $\pi_j(1) > .9$ and $\pi_j(0) < .1$, and our sample size recommendations (Sec. 6) will prescribe performing the MSA experiment with a sample in which good and defective specimens have about equal presence (i.e., $\theta \approx .5$).
- Parameter estimates of both methods become heavily biased if the identifiability restrictions (2) are close to being violated by all parameters, for example, θ being close to 1.00 and the $\pi_j(1)$'s being only slightly larger than the $\pi_j(0)$'s. To illustrate this, consider the one-rater situation with parameters $(\theta, \pi_1(1), \pi_1(0)) = (.90, .55, .40)$ and $(n, \ell) = (150, 15)$. The average of estimates $(\hat{\theta}, \hat{\pi}_1(1), \hat{\pi}_1(0))$ yielded by simulations are $(.638, .603, .385)$ for ML and $(.726, .571, .422)$ for MoM.
- When the experiment involves poor inspection systems $[1 \gg \pi_j(1) > \pi_j(0) \gg 0]$ and the objects used in the MSA experiment are far from balanced over good and defective (say $\theta \approx .95$), the ML method produces estimates that are less biased and have a smaller standard deviation than the MoM.
- The MoM sometimes produces estimates that are outside the parameter space, for example, $\hat{\pi}_1(1) > 1.00$. The frequency at which this occurs depends on how close parameters are to violating the identifiability restrictions. It also depends on the sample size; increasing n and ℓ decreases this frequency. The problem is unlikely to occur if $\theta \approx .5$ and $\pi_j(1) \gg \pi_j(0)$.

A sample of the simulation results is given in Table A.1 in the Appendix. Given the results of the simulation study, we recommend performing the MSA experiment with a sample of objects in which good and defective specimens are approximately equally represented. The simulation showed that in the practically most relevant situation, where the inspection systems have good specificity and sensitivity and the quality of the sample is balanced, both methods can be used. The closed expression estimators of the MoM may even be preferred over the algorithmic approach of the ML. Alternatively, the MoM estimates could be used as an initial guess in the EM algorithm, decreasing the number of iterations to satisfy the stopping criterion. But when either the specificity and sensitivity are poor or the sample is nonbalanced, the ML method is preferred, although both methods then produce biased and highly variable estimates.

5. MODEL DIAGNOSTICS

5.1 Goodness of Fit

When fitting the latent class model to the experimental data, it is important to assess the validity of the model. The issue of model checking when using latent class models has been addressed in the biostatistical literature (Collins, Fidler, Wugalter, and Long 1993; Garrett and Zeger 2000; Formann 2003a,b). The need for model checking becomes apparent from the possible biasedness of the estimates when the identifiability restrictions are close to being violated, or when the conditional independence assumption does not hold (Torrance-Rynard and Walter 1997).

A commonly used goodness-of-fit test for latent class models, proposed by Collins, Fidler, Wugalter, and Long (1993), evaluates the null hypothesis $H_0 : \boldsymbol{\Psi} = \hat{\boldsymbol{\Psi}}$ by comparing the response frequencies predicted by the model with the observed response frequencies. Their test uses a special case ($\lambda = \frac{2}{3}$) of the power-divergence statistic of Cressie and Read (1984), which is given by

$$\frac{2}{\lambda(\lambda+1)} \sum_{\mathbf{R} \in \{0,\ldots,\ell\}^m} \{\#i : 1 \le i \le n, \mathbf{r}_i = \mathbf{R}\} \times \left[\left(\frac{\{\#i : 1 \le i \le n, \mathbf{r}_i = \mathbf{R}\}}{\mathrm{E}_{\hat{\boldsymbol{\Psi}}}(\{\#i : 1 \le i \le n, \mathbf{r}_i = \mathbf{R}\})} \right)^{\lambda} - 1 \right]. \quad (7)$$

It sums over response patterns $\mathbf{R}_i = (R_{i1}, \ldots, R_{im})$, and for the observed response patterns accumulates the extent of dissimilarity between observed and expected frequencies. This is a generalization of Pearson's chi-squared statistic ($\lambda = 1$) and the log-likelihood statistic G^2 ($\lambda = 0$). It is approximately chi-squared distributed with $(\ell+1)^m - 1 - (2m+1)$ degrees of freedom (the total number of different responses minus 1, and one subtracted for each parameter estimated). Cressie and Read (1984) pointed out that this statistic with $\lambda = \frac{2}{3}$ is less sensitive to low frequencies than Pearson's chi-squared statistic and the log-likelihood statistic G^2. But Formann (2003a) pointed out that if $\lambda = 0, \frac{2}{3}$ or 1, then problems arise for sparse data. A better choice then may be $\lambda = -\frac{1}{2}$, resulting in the Freeman–Tukey statistic. This may be an advantage, especially if the R&R are good. In that case, response patterns with disagreement may have low frequencies.

The chi-squared approximation to the distribution of the goodness-of-fit test statistic (7) may be poor for sparse data and when expected frequencies are small (both of which occur in situations with near-perfect sensitivity and specificity). The null distribution of the test is then better obtained empirically through Monte Carlo resampling (also called the parametric bootstrap). Toward this end, we draw new experimental data from model (1) with estimated $\hat{\boldsymbol{\Psi}}$ and reestimate the parameters from the new data. The test statistic (7) is calculated from the new data and reestimated parameters. This process (resampling, estimation, and calculation) is repeated for $c = 1, \ldots, M$. Let T_{obs} denote the observed test statistic and let T_c, $c = 1, \ldots, M$, denote the resampled test statistics. The p value for the goodness-of-fit test is calculated as

$$\{\#c \le M : T_{\mathrm{obs}} < T_c\}/M.$$

One option for residual analysis is the standardized residuals. Formann (2003a) gave alternative definitions for residuals in categorical data, particularly the Freeman–Tukey variance-stabilized residuals, which can be best used for sparse data. They are defined as

$$\sqrt{\{\#i : \mathbf{R}_i = \mathbf{r}\}} + \sqrt{\{\#i : \mathbf{R}_i = \mathbf{r}\} + 1} - \sqrt{4\mathrm{E}_{\hat{\boldsymbol{\Psi}}}(\{\#i : \mathbf{R}_i = \mathbf{r}\}) + 1}. \quad (8)$$

Finally, together with the parameter estimates, the confidence intervals of the estimates should be reported (De Menezes 1999).

5.2 Reproducibility

In the ML estimation framework, we can test for the presence of reproducibility issues. The likelihood ratio test statistic is defined by

$$2(\mathcal{L}_{\mathrm{RR}} - \mathcal{L}_{\mathrm{R}}), \tag{9}$$

where $\mathcal{L}_{\mathrm{RR}}$ is the log-likelihood of the model as described in Section 2 (with both objects and raters effects, accommodating repeatability and reproducibility). $\mathcal{L}_{\mathrm{R}}$ is the model with objects effects only (accommodating repeatability but assuming no rater effects). In the latter model, $\pi_1(0) = \cdots = \pi_m(0)$ and $\pi_1(1) = \cdots = \pi_m(1)$. The test statistic, evaluated at the corresponding ML estimates for the parameters, is approximately chi-squared distributed with $2m + 1 - 3$ degrees of freedom.

By comparing deviances (McCullagh and Nelder 1989), we may quantify the proportions of repeatability and reproducibility. We propose

$$D_{\text{repeatability}} = 2(\mathcal{L}_{Saturated} - \mathcal{L}_{\mathrm{RR}}) \tag{10}$$

and

$$D_{\text{reproducibility}} = 2(\mathcal{L}_{\mathrm{RR}} - \mathcal{L}_{\mathrm{R}}). \tag{11}$$

The saturated model referred to is a model in which all observed variation is accounted for by the systematic part of the model. In models based on continuous distributions, the saturated model has as many parameters as there are observations; the corresponding log-likelihood typically is smaller than 0. In the ANOVA type of analysis of standard-gauge R&R experiments, these deviances reduce to multiples of sums of squares. In our case the saturated model is not $R_{ij} \sim \mathrm{B}(\ell; p_{ij})$ (binomial with parameters ℓ and p_{ij}), which was considered by McCullagh and Nelder (1989, pp. 188 ff.), because this model still treats a part of the observed variation as stochastic, and, consequently, (10) would represent only part of the repeatability. Instead, the saturated model is

$$X_{ijh} \sim \mathrm{B}(1, p_{ijh}), \quad \text{with ML fit } \hat{p}_{ijh} = X_{ijh}.$$

Consequently, $\mathcal{L}_{Saturated} = 0$, and (10) reduces to $D_{\text{repeatability}} = -2\mathcal{L}_{\mathrm{RR}}$.

6. SAMPLE SIZE

Our sample size recommendations include the advice to attempt to select a sample of objects in which good and defective objects have about equal representation. We do not assume that a gold standard is available (i.e., that the objects' true values are known). When a gold standard is available, the ML estimators reduce to simple proportions, and our methods are not needed (Automotive Industry Action Group 2002, pp. 125–140, Danila, Steiner, and Mackay 2008). But even if reference values are unknown, it often is possible for the user to influence the proportion θ in the sample, typically because he or she can sample from two subpopulations (e.g., the streams of approved and rejected objects), one with θ near 0 and the other with θ near 1. Our method does not require that the user can do so, but if the user can, then he or she can obtain a sample with θ closer to .5 than in the original population. This will improve the reliability of his MSA study, because our method's precision is better as θ is closer to .5.

To give sample size recommendations, we conducted a second simulation. In line with the foregoing, we assume that θ is close to .5 and draw the θ from Beta(50, 50). Furthermore, we assume that the sensitivity, $\pi_j(1)$, is closer to 1 than .8 and the specificity, $1 - \pi_j(0)$, is closer to 1 than .8. The $\pi_j(1)$'s are drawn from Beta(25, 1), and the $\pi_j(0)$'s are drawn from Beta(1, 25). For each iteration of the simulation, new realizations of θ, $\pi_j(1)$, and $\pi_j(0)$ are drawn, and the data, $\mathbf{X}$, are drawn in accordance with distribution (1). The number of objects n and repetitions ℓ varied from 25 to 150 and 1 (identifiability restrictions allowing) to 11. For each combination, the standard error of the estimates is calculated on the basis of 10,000 iterations as

$$\mathrm{se}(\pi_j(1)) = \sqrt{\frac{1}{10{,}000}\sum_{s=1}^{10{,}000}\left(\hat{\pi}_j^{(s)}(1) - \pi_j^{(s)}(1)\right)^2},$$

where $\theta^{(s)}$, $\pi_j^{(s)}(1)$, and $\pi_j^{(s)}(0)$ are the drawn parameters for the sth iteration and $\hat{\theta}^{(s)}$, $\hat{\pi}_j^{(s)}(1)$, and $\hat{\pi}_j^{(s)}(0)$ are their respective ML estimates. This is done for $m = 1$–4 raters.

The results are summarized in Figure 1, where the standard errors of all $\pi_j(1)$ and $\pi_j(0)$ are averaged, because they stem from the same distribution. The results of θ are omitted, because θ is not relevant for evaluating the measurement system. From Figure 1, we can see that (for two or more raters), much precision is gained by designing the experiment with $\ell = 3$ instead of $\ell = 2$ repetitions. (For $m = 1$ rater, ℓ should be at least 3, to ensure identifiability of the model.) Including 50 (60 in the case of 1 rater) objects in the experiment ensures that twice the standard error is smaller than .05. Our recommendations can be summarized as follows:

- Obtain a sample of objects in which good and defective specimens have about equal shares.
- A sample size of 50 (in the case of 1 rater, 60) objects is sufficient.
- Because the factor related to raters is treated as a fixed effect in the model, the choice of the number of raters will be driven more by pragmatic considerations than statistical considerations.
- Design the experiment to include three repetitions per rater per object.
- Make sure that the suspected good and defective specimens are random samples from the subpopulations of good and defective objects.

The final recommendation ensures conditional exchangeability; that is, given the Y_i, the objects are exchangeable with respect to X_{ijh} (see Lindley and Novick 1981). Then de Finetti's theorem allows us to extrapolate the estimated $\hat{\pi}_i(\cdot)$ to the population of all objects, even if the conditional independence assumption does not hold (again see Lindley and Novick 1981).

7. EXAMPLE

In a real-life example, we demonstrate our methods and also show the typical problems that users may run into and demonstrate how our methods provide warning diagnostics and a certain degree of robustness. We consider the outgoing inspection of an injection molding process. The company applies an

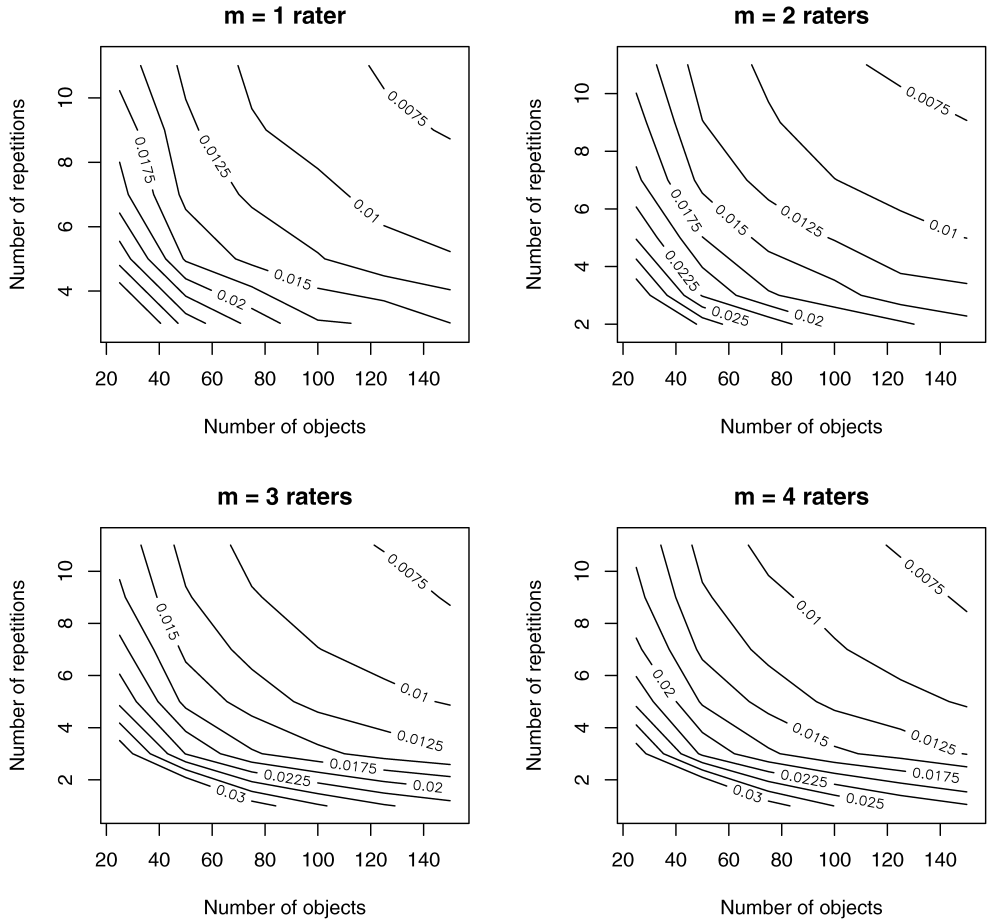

Figure 1. Contour plots of the standard error of the estimates of the $\pi_j(\cdot)$ for 1 to 4 raters.

inspection plan based on AQL and LQL specifications. The molded parts are visually inspected for imperfections and defects, such as splay marks, short shots, broken tabs, scratches, and so on. Parts are rated as "accept" or "reject," and, based on the number of rejected parts in the sample, the lot is either rejected or shipped to the customer.

The current inspection scheme was not satisfactory for both the customer and the company itself in terms of inspection efforts and the number of lots of substandard quality reaching the customer. As a first step in designing a better inspection process, an MSA experiment was done to determine the R&R of the visual inspections. The process engineer collected 80 parts, deliberately aiming for a sample with about equal proportions of good and defective specimens. He did so by taking a random sample of about (but not precisely) 40 items from a container of scrapped parts and a second sample of about 40 items from the stream of accepted parts. Each part was judged twice by three operators. Note that the setup of this experiment deviates from the recommendations given in the previous section; a setup with 50 parts and 3 repetitions per operator would give a slightly better precision (see Fig. 1) at a slightly smaller sample size.

The results of the experiment are summarized in the form of response patterns (R_{i1}, R_{i2}, R_{i3}). There are 27 potential response patterns, only 22 of which occurred in the experiment. Table 1 gives the results.

The ML and MoM estimators of the model parameters, along with their 95% bootstrapped confidence intervals (CIs), are given in Table 2. The table readily translates into an intrarater analysis (i.e., repeatability). The sensitivities of the three operators (with ML estimates .75, .79, and .81) are fair. The specificities are good for operators 1 and 2 but substantially poorer for operator 3 $[1 - \hat{\pi}_j(0) = .92, .97, \text{and } .69]$. Note how the results presented in Table 2 give tangible directions for identifying the problem with the inspection procedure. The user can verify whether the problem is in the sensitivity or specificity, and also can evaluate whether the problem is similar for all raters or more pronounced for a single rater.

Provided that an estimate $\hat{\theta}_p$ of the process's true defect rate is available, the per-rater probabilities of misclassification are given as $\hat{\theta}_p(1 - \hat{\pi}_j(1)) + (1 - \hat{\theta}_p)\hat{\pi}_j(0)$. The estimated θ reported in the table cannot be used, because it represents the proportion of defects in the sample, not in the process. Note that the CIs given in Table 2 are larger than would be expected from the simulated standard errors reported in Figure 1. This can be explained by the different parameter values used in the simulation; drawing from Beta distributions with means equal to the estimated parameters yields standard errors in agreement with the observed confidence intervals.

We apply the likelihood ratio test (9) for reproducibility issues. The fitted parameters of the model with no reproducibility issues are $\hat{\theta} = .39$, $\hat{\pi}(1) = .80$, and $\hat{\pi}(0) = .15$ (ML estimates). The test statistic equals $2(-215.75 + 230.31) = 29.12$, which, based on the chi-squared approximation with $7 - 3 = 4$ degrees of freedom, corresponds to a p value $< .001$, confirming that

Table 1. Frequency table of the response patterns with corresponding Freeman–Tukey residuals

Response pattern	Observed frequency	Expected frequency, ML	Freeman–Tukey residual, ML	Expected frequency, MoM	Freeman–Tukey residual, MoM
(0, 0, 0)	22	18.12	.91	19.39	.62
(0, 0, 1)	12	16.12	−1.02	14.36	−.58
(0, 0, 2)	4	3.63	.30	2.74	.78
(0, 1, 0)	1	.96	.22	3.09	−1.24
(0, 1, 1)	1	1.04	.14	2.43	−.86
(0, 1, 2)	0	.63	−.88	.87	−1.12
(0, 2, 0)	1	.06	1.30	.14	1.17
(0, 2, 1)	1	.40	.80	.28	.96
(0, 2, 2)	2	.83	1.07	.56	1.35
(1, 0, 0)	1	3.27	−1.34	1.93	−.54
(1, 0, 1)	4	3.06	.60	1.65	1.48
(1, 0, 2)	3	1.00	1.49	.96	1.54
(1, 1, 0)	0	.32	−.51	.40	−.62
(1, 1, 1)	1	1.43	−.18	1.39	−.15
(1, 1, 2)	0	2.72	−2.45	3.39	−2.81
(1, 2, 0)	1	.28	.95	.13	1.17
(1, 2, 1)	1	2.37	−.82	1.41	−.16
(1, 2, 2)	4	4.96	−.33	4.03	.10
(2, 0, 0)	0	.18	−.31	.09	−.16
(2, 0, 1)	1	.39	.81	.49	.70
(2, 0, 2)	3	.58	1.91	1.30	1.24
(2, 1, 0)	0	.24	−.40	.20	−.34
(2, 1, 1)	2	1.93	.19	2.18	.03
(2, 1, 2)	1	4.04	−1.73	6.24	−2.68
(2, 2, 0)	1	.42	.77	.23	1.03
(2, 2, 1)	1	3.56	−1.49	2.62	−.97
(2, 2, 2)	12	7.46	1.52	7.51	1.50
Total	80				

there are reproducibility issues. This means that the consistency across raters is significantly worse than the average intra-rater consistency. Note that repeatability is the larger contributor in deviance; the $480-7=473$ degrees of freedom associated with repeatability account for a deviance of $D_{\text{repeatability}} = 431.5$, whereas the $7-3=4$ degrees of freedom associated with the raters effects account for a deviance of $D_{\text{reproducibility}} = 29.12$.

For the current inspection system, we find an estimated total sensitivity of $Se = (.75 + .79 + .81)/3 = .78$ and specificity of $Sp = (.92 + .97 + .69)/3 = .86$ (ML estimates). The misclassification probability is $\hat{\theta}_p(1 - .78) + (1 - \hat{\theta}_p)(1 - .86) = .08\hat{\theta}_p + .14$.

As mentioned earlier, goodness-of-fit checking is important in latent class modeling, and we address the matter in some length. The first question is whether the method used, with its assumed homogeneity of the subpopulations of good and defective objects, is likely to be valid on a priori grounds. We expect that the subpopulation of good objects (having no imperfections) can be considered homogeneous, but we are less confident about the subpopulation of defective objects. Imperfections can be splay marks, short shots, broken tabs, scratches, and others, and although this is not expected, it may be the case that not all of these are detected with identical R&R, thus violating the conditional independence assumption.

Table 2. ML and MoM estimates and corresponding 95% CIs of the parameters

	ML	95% CI_{ML}	MoM	95% CI_{MoM}
θ	.41	(.27; .53)	.42	(.29; .58)
$\pi_1(1)$	.75	(.61; .91)	.79	(.62; .95)
$\pi_2(1)$	.79	(.59; 1.00)	.71	(.53; .87)
$\pi_3(1)$	.81	(.68; .95)	.85	(.70; .97)
$\pi_1(0)$	.08	(.01; .21)	.05	(0; .12)
$\pi_2(0)$	.03	(0; .12)	.07	(0; .16)
$\pi_3(0)$	.31	(.18; .44)	.27	(.16; .38)

The goodness-of-fit test statistic (7) with $\lambda = -\frac{1}{2}$ equals $C_{ML} = 43.80$ and $C_{MoM} = 51.08$ for the ML and MoM estimates. The corresponding p values based on the Monte Carlo resampling are .065 and .025, leading to a rejection for the MoM fit. We examine the residuals in Table 1 for clues, focusing on the ML estimates and looking for patterns in the large deviations between observed and expected frequencies (or, alternatively, large FT residuals). The extreme response patterns (0, 0, 0) and (2, 2, 2) are overrepresented at the expense of such patterns as (0, 0, 1) and (1, 0, 0) or (2, 2, 1) and (2, 1, 2). Furthermore, response patterns representing extreme disagreement among the raters tend to be overrepresented; for example, the response patterns that have at least one 0 and one 2, such as (2, 0, 2), have a total observed frequency of 17 versus a total expected frequency of 8.64. One way to interpret this is that the sample of objects has a substantial number of extremely good

and extremely bad specimens that are rated with very high sensitivity and specificity and a substantial number of specimens that are in a gray area in between good and defective, creating disagreement among the raters. Our model cannot accommodate this, because it assumes the same specificity and sensitivity for all good and defective objects (i.e., there are no extremely good or extremely bad items, nor is there a gray area in between). So in effect, the large residuals hint at a violation of the conditional independence assumption (which states that sensitivity and specificity do not depend on the degree of badness or goodness but are identical for all good or defective objects). In summary, the goodness-of-fit test and residual analysis indicate that the MoM fit should be rejected and provide some (albeit not significant) evidence that the conditional independence assumption does not hold.

In the event that the conditional independence assumption is rejected, the user still can use the estimated $\pi_j(y)$ and probabilities of misclassification, provided that the sampling method ensures conditional exchangeability of objects (boiling down to random sampling from both subpopulations, as explained in the previous section). We have doubts here, because it is plausible that objects with minor imperfections may be somewhat underrepresented in the container from which supposedly defective objects were sampled, because they are more likely to slip through the inspection procedure than objects with major imperfections. If the conditional independence assumption does not hold, then this would result in a slight overestimation of sensitivity and specificity.

8. CONCLUSION

The standard methodology for standard-gauge R&R studies cannot be applied in the case of measurements on a binary scale, such as pass/fail inspection. If it is possible to determine the true values of the objects in the sample (by some authoritative inspection system), then determining sensitivity and specificity is fairly straightforward. We focus on the more challenging situation in which this gold standard is not available. Unlike current accounts, our experimental setup involves both an objects factor and a raters factor, as well as replications. Under a favorable scenario, we showed that such an experiment would require a sample size of $n = 50$ objects that are measured $\ell = 3$ times by each rater. Along with ML estimators, we outline how to obtain closed-expression estimators based on the MoM. The example illustrates that model adequacy checking is important, and the proposed approach provides a goodness-of-fit test and residual analysis.

The purpose of this article was to extend and refine the methodology proposed by Boyles (2001) and in the biostatistical literature. Essential to this methodology is that the underlying true ("reference") value be represented as a dichotomy, such as good versus bad or functional versus defective. Binary measurements also can be regarded as a limiting case of ordinal measurements. In that case, one would represent the underlying characteristic as a continuum and use a latent trait model instead of a latent class model (see De Mast and Van Wieringen 2008). Finally, one can consider binary rating as a limiting case of nominal measurement and adopt methods based on the agreement concept (De Mast and Van Wieringen 2007). All of these approaches are based on latent class modeling; the differences are in the details of the models, the parameter estimation methods, and the metrics used for expressing R&R. In future work, we intend to study how these three options compare, aiming to arrive at recommendations for their use for binary ratings.

ACKNOWLEDGMENTS

The authors thank the editor and anonymous associate editor for their constructive comments on early versions of this work that led to substantial improvements in the article.

APPENDIX: COMPARISON SIMULATION RESULTS

Table A.1 additional material for Section 4.

Table A.1. Sample of the results of the simulation study comparing the MoM and ML estimators

Scenario	m	Method	n	ℓ	θ	$\bar{\hat{\theta}}$	se($\hat{\theta}$)	$\pi_1(1)$	$\bar{\hat{\pi}}_1(1)$	se($\hat{\pi}_1(1)$)	$\pi_1(0)$	$\bar{\hat{\pi}}_1(0)$	se($\hat{\pi}_1(0)$)
1	3	MoM	20	2	.9	.769	.123	.95	.971	.034	.40	.492	.231
2	3	MoM	20	10	.9	.868	.067	.95	.952	.018	.40	.409	.153
3	3	MoM	60	6	.9	.889	.042	.95	.952	.014	.40	.399	.139
4	3	MoM	100	2	.9	.850	.074	.95	.958	.021	.40	.455	.187
5	3	MoM	100	10	.9	.899	.031	.95	.950	.008	.40	.398	.082
1	3	ML	20	2	.9	.800	.176	.95	.953	.056	.40	.508	.333
2	3	ML	20	10	.9	.887	.089	.95	.951	.018	.40	.467	.209
3	3	ML	60	6	.9	.898	.041	.95	.950	.012	.40	.402	.099
4	3	ML	100	2	.9	.883	.072	.95	.952	.021	.40	.401	.190
5	3	ML	100	10	.9	.899	.030	.95	.950	.007	.40	.401	.052
6	3	MoM	20	2	.5	.515	.130	.95	.932	.060	.40	.380	.145
7	3	MoM	20	10	.5	.501	.112	.95	.949	.026	.40	.398	.057
8	3	MoM	60	6	.5	.501	.070	.95	.950	.021	.40	.399	.044
9	3	MoM	100	2	.5	.509	.078	.95	.943	.034	.40	.392	.079
10	3	MoM	100	10	.5	.499	.052	.95	.951	.012	.40	.400	.025

Table A.1. (Continued)

Scenario	m	Method	n	ℓ	θ	$\bar{\hat{\theta}}$	se($\hat{\theta}$)	$\pi_1(1)$	$\bar{\hat{\pi}}_1(1)$	se($\hat{\pi}_1(1)$)	$\pi_1(0)$	$\bar{\hat{\pi}}_1(0)$	se($\hat{\pi}_1(0)$)
6	3	ML	20	2	.5	.512	.152	.95	.942	.072	.40	.382	.154
7	3	ML	20	10	.5	.499	.110	.95	.950	.023	.40	.399	.050
8	3	ML	60	6	.5	.500	.066	.95	.950	.017	.40	.400	.038
9	3	ML	100	2	.5	.500	.070	.95	.951	.032	.40	.398	.064
10	3	ML	100	10	.5	.499	.050	.95	.950	.010	.40	.400	.022
11	3	MoM	20	2	.9	.754	.126	.95	.961	.041	.10	.432	.232
12	3	MoM	20	10	.9	.835	.089	.95	.951	.024	.10	.294	.204
13	3	MoM	60	6	.9	.854	.065	.95	.952	.020	.10	.296	.178
14	3	MoM	100	2	.9	.821	.088	.95	.951	.026	.10	.396	.197
15	3	MoM	100	10	.9	.879	.041	.95	.950	.014	.10	.209	.133
11	3	ML	20	2	.9	.829	.149	.95	.953	.066	.10	.279	.343
12	3	ML	20	10	.9	.890	.082	.95	.950	.017	.10	.203	.281
13	3	ML	60	6	.9	.899	.040	.95	.950	.010	.10	.104	.069
14	3	ML	100	2	.9	.891	.047	.95	.953	.022	.10	.121	.132
15	3	ML	100	10	.9	.900	.030	.95	.950	.010	.10	.100	.032
16	3	MoM	20	2	.5	.511	.127	.95	.861	.099	.10	.177	.115
17	3	MoM	20	10	.5	.506	.110	.95	.916	.058	.10	.122	.068
18	3	MoM	60	6	.5	.509	.077	.95	.920	.052	.10	.116	.063
19	3	MoM	100	2	.5	.508	.085	.95	.883	.072	.10	.152	.084
20	3	MoM	100	10	.5	.505	.056	.95	.941	.031	.10	.101	.039
16	3	ML	20	2	.5	.507	.126	.95	.941	.075	.10	.096	.094
17	3	ML	20	10	.5	.499	.110	.95	.950	.022	.10	.100	.031
18	3	ML	60	6	.5	.500	.065	.95	.950	.017	.10	.100	.023
19	3	ML	100	2	.5	.500	.058	.95	.950	.033	.10	.100	.042
20	3	ML	100	10	.5	.499	.050	.95	.950	.010	.10	.100	.014

Scenario	Method	$\pi_2(1)$	$\bar{\hat{\pi}}_2(1)$	sd($\hat{\pi}_2(1)$)	$\pi_3(1)$	$\bar{\hat{\pi}}_3(1)$	sd($\hat{\pi}_3(1)$)	$\pi_2(0)$	$\bar{\hat{\pi}}_2(0)$	sd($\hat{\pi}_2(0)$)	$\pi_3(0)$	$\bar{\hat{\pi}}_3(0)$	sd($\hat{\pi}_3(0)$)	Realizations
1	MoM	.75	.798	.085	.550	.610	.101	.25	.264	.187	.10	.131	.129	1,597
2	MoM	.75	.753	.035	.550	.554	.039	.25	.255	.113	.10	.115	.076	7,018
3	MoM	.75	.753	.027	.550	.553	.030	.25	.249	.094	.10	.109	.061	8,348
4	MoM	.75	.769	.043	.550	.568	.049	.25	.272	.134	.10	.138	.089	4,810
5	MoM	.75	.750	.015	.550	.550	.017	.25	.249	.055	.10	.101	.037	9,871
1	ML	.75	.774	.104	.550	.587	.125	.25	.316	.290	.10	.159	.218	10,000
2	ML	.75	.751	.035	.550	.553	.041	.25	.303	.179	.10	.143	.144	10,000
3	ML	.75	.751	.024	.550	.551	.028	.25	.252	.086	.10	.102	.061	10,000
4	ML	.75	.755	.040	.550	.556	.045	.25	.252	.156	.10	.107	.109	10,000
5	ML	.75	.750	.014	.550	.550	.017	.25	.251	.046	.10	.100	.032	10,000
6	MoM	.75	.771	.112	.550	.580	.135	.25	.221	.111	.10	.081	.067	4,019
7	MoM	.75	.751	.048	.550	.552	.055	.25	.249	.047	.10	.100	.033	9,709
8	MoM	.75	.751	.038	.550	.551	.042	.25	.249	.037	.10	.100	.026	9,904
9	MoM	.75	.752	.062	.550	.552	.068	.25	.241	.059	.10	.094	.040	8,556
10	MoM	.75	.750	.021	.550	.550	.024	.25	.250	.021	.10	.100	.015	9,999
6	ML	.75	.755	.127	.550	.559	.147	.25	.236	.127	.10	.093	.086	10,000
7	ML	.75	.751	.045	.550	.550	.052	.25	.250	.044	.10	.100	.030	10,000
8	ML	.75	.750	.033	.550	.550	.038	.25	.249	.033	.10	.100	.023	10,000
9	ML	.75	.751	.055	.550	.551	.062	.25	.249	.054	.10	.099	.039	10,000
10	ML	.75	.750	.020	.550	.550	.023	.25	.250	.019	.10	.100	.013	10,000
11	MoM	.75	.817	.087	.550	.628	.093	.25	.266	.189	.40	.238	.159	1,032
12	MoM	.75	.775	.047	.550	.564	.039	.25	.260	.128	.40	.354	.094	3,837
13	MoM	.75	.774	.037	.550	.560	.028	.25	.249	.111	.40	.366	.077	4,275
14	MoM	.75	.796	.055	.550	.578	.042	.25	.245	.130	.40	.322	.102	2,347
15	MoM	.75	.760	.020	.550	.554	.017	.25	.244	.078	.40	.394	.050	6,129
11	ML	.75	.769	.093	.550	.571	.113	.25	.293	.273	.40	.380	.277	10,000
12	ML	.75	.751	.035	.550	.552	.040	.25	.301	.174	.40	.407	.126	10,000
13	ML	.75	.750	.024	.550	.551	.028	.25	.251	.083	.40	.398	.091	10,000
14	ML	.75	.753	.035	.550	.552	.039	.25	.251	.124	.40	.399	.131	10,000
15	ML	.75	.750	.014	.550	.550	.017	.25	.250	.046	.40	.401	.051	10,000

Table A.1. (Continued)

Scenario	Method	$\pi_2(1)$	$\bar{\hat{\pi}}_2(1)$	$\text{sd}(\hat{\pi}_2(1))$	$\pi_3(1)$	$\bar{\hat{\pi}}_3(1)$	$\text{sd}(\hat{\pi}_3(1))$	$\pi_2(0)$	$\bar{\hat{\pi}}_2(0)$	$\text{sd}(\hat{\pi}_2(0))$	$\pi_3(0)$	$\bar{\hat{\pi}}_3(0)$	$\text{sd}(\hat{\pi}_3(0))$	Realizations
16	MoM	.75	.803	.110	.550	.644	.110	.25	.187	.108	.40	.298	.110	1,690
17	MoM	.75	.768	.061	.550	.563	.049	.25	.226	.057	.40	.385	.047	5,828
18	MoM	.75	.766	.053	.550	.558	.037	.25	.226	.049	.40	.389	.037	6,083
19	MoM	.75	.795	.077	.550	.576	.051	.25	.194	.073	.40	.370	.051	3,394
20	MoM	.75	.752	.030	.550	.552	.023	.25	.243	.027	.40	.398	.022	8,845
16	ML	.75	.755	.113	.550	.554	.123	.25	.238	.111	.40	.396	.126	10,000
17	ML	.75	.751	.044	.550	.551	.051	.25	.250	.044	.40	.400	.050	10,000
18	ML	.75	.750	.033	.550	.550	.037	.25	.249	.033	.40	.400	.037	10,000
19	ML	.75	.751	.048	.550	.552	.053	.25	.249	.050	.40	.399	.052	10,000
20	ML	.75	.750	.019	.550	.550	.023	.25	.250	.019	.40	.400	.022	10,000

NOTE: This table gives the results (average and standard errors) in the $m = 3$ raters situation for various choices of number of objects (n) and replications (ℓ), and various choices of the model parameters. The results are based on a simulation involving 10,000 or less realizations.

[Received October 2006. Revised July 2008.]

REFERENCES

Automotive Industry Action Group (2002), *Measurement System Analysis: Reference Manual* (3rd ed.), Detroit, MI: Author.

Allen, M. J., and Yen, W. M. (1979), *Measurement Theory*, Monterey, CA: Brooks/Cole.

Bartholomew, D. J., and Knott, M. (1999), *Latent Variable Models and Factor Analysis*, New York: Oxford University Press.

Blischke, W. R. (1962), "Moment Estimators for the Parameters of a Mixture of Two Binomial Distributions," *Annals of Mathematical Statistics*, 33, 444–454.

Boyles, R. A. (2001), "Gauge Capability for Pass–Fail Inspection," *Technometrics*, 43, 223–229.

Burdick, R. K., Borror, C. M., and Montgomery, D. C. (2003), "A Review of Methods for Measurement Systems Capability Analysis," *Journal of Quality Technology*, 35, 342–354.

Collins, L. M., Fidler, P. L., Wugalter, S. E., and Long, J. D. (1993), "Goodness-of-Fit Testing for Latent Class Models," *Multivariate Behavioral Research*, 28, 375–389.

Cressie, N., and Read, T. R. C. (1984), "Multinomial Goodness-of-Fit Tests," *Journal of the Royal Statistical Society*, Ser. B, 46, 440–464.

Danila, O., Steiner, S. H., and Mackay, R. J. (2008), "Assessing a Binary Measurement System," *Journal of Quality Technology*, 40, 310–318.

De Mast, J., and Van Wieringen, W. N. (2007), "Measurement System Analysis for Categorical Measurements: Agreement and Kappa-Type Indices," *Journal of Quality Technology*, 39, 191–202.

——— (2008), "Modeling and Evaluating Repeatability and Reproducibility of Ordinal Classifications," unpublished manuscript.

De Menezes, L. M. (1999), "On Fitting Latent Class Models for Binary Data: The Estimation of Standard Errors," *British Journal of Mathematical and Statistical Psychology*, 52, 149–168.

Formann, A. K. (2003a), "Latent Class Model Diagnostics: A Review and Some Proposals," *Computational Statistics & Data Analysis*, 41, 549–559.

——— (2003b), "Latent Class Model Diagnostics From a Frequentist Point of View," *Biometrics*, 59, 189–196.

Garrett, E. S., and Zeger, S. (2000), "Latent Class Model Diagnosis," *Biometrics*, 56, 1055–1067.

Garrett, E. S., Eaton, W. W., and Zeger, S. (2002), "Methods for Evaluating the Performance of Diagnostic Tests in the Absence of a Gold Standard: A Latent Class Model Approach," *Statistics in Medicine*, 21, 1289–1307.

Hui, S. L., and Walter, S. D. (1980), "Estimating the Error Rates of Diagnostic Tests," *Biometrics*, 36, 167–171.

Lindley, D. V., and Novick, M. R. (1981), "The Role of Exchangeability in Inference," *The Annals of Statistics*, 9, 45–58.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman & Hall.

McLachlan, G. J., and Krishnan, T. (1997), *The EM Algorithm and Extensions*, New York: Wiley.

Qu, Y., Tan, T., and Kutner, M. H. (1996), "Random Effects Models in Latent Class Analysis for Evaluating Accuracy of Diagnostic Tests," *Biometrics*, 52, 797–810.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: Wiley.

Torrance-Rynard, V. L., and Walter, S. D. (1997), "Effects of Dependent Errors in the Assessment of Diagnostic Test Performance," *Statistics in Medicine*, 16, 2157–2175.

Van Wieringen, W. N. (2003), "Statistical Models for the Precision of Categorical Measurement Systems," unpublished doctoral thesis, University of Amsterdam, Dept. of Mathematics.

——— (2005), "On Identifiability of Certain Latent Class Models," *Statistics and Probability Letters*, 75, 211–218.

Van Wieringen, W. N., and Van den Heuvel, E. R. (2005), "A Comparison of Methods for the Evaluation of Binary Measurement Systems," *Quality Engineering*, 17, 495–507.

Vardeman, S. B., and Van Valkenburg, E. S. (1999), "Two-Way Random-Effects Analyses and Gauge R&R Studies," *Technometrics*, 41, 202–211.