*The Canadian Journal of Statistics*
Vol. 39, No. 2, 2011, Pages 344–355
*La revue canadienne de statistique*

# Planning and analysis of measurement reliability studies

Stefan H. STEINER[1]*, Nathaniel T. STEVENS[1], Ryan BROWNE[2] and R. Jock MACKAY[1]

[1] *Business and Industrial Statistics Research Group, Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, Canada N2L 3G1*
[2] *Department of Mathematics and Statistics, University of Guelph, Guelph, Canada N1G 2W1*

*Key words and phrases:* Gauge repeatability; gauge reproducibility; intraclass correlation; random effects; study design

*MSC 2010:* Primary 62P30.

*Abstract:* In the traditional plan for assessing the reliability of a measurement system, a number of raters each measure the same group of subjects. If the system has a large number of raters, we recommend a new set of plans that has two advantages over the traditional plan. First, the proposed plans provide greater precision for estimating the intraclass correlation coefficient with the same total number of measurements. Second, the plans are flexible and can be adapted to constraints on the number of times any subject can be assessed or the number of times any rater can make an assessment. We provide a simple tool for planning a reliability study, access to the software for the planning in the case where there are constraints and an example to demonstrate the analysis of data from the proposed plans. *The Canadian Journal of Statistics* 39: 344–355; 2011 © 2011 Statistical Society of Canada

*Résumé:* Dans un plan traditionnel pour déterminer la fiabilité d'un système de mesures, plusieurs évaluateurs mesurent tous les sujets d'un même groupe. Lorsqu'il y a un grand nombre d'évaluateurs, nous recommandons un nouvel ensemble de plans qui possède deux avantages par rapport au plan traditionnel. Premièrement, les plans proposés procurent une plus grande précision pour l'estimation du coefficient de corrélation intraclasse avec un même nombre de mesures. Deuxièmement, ces plans sont flexibles et ils peuvent être modifiés pour contraindre le nombre d'évaluations par sujet ou encore le nombre de mesures faites par un évaluateur. Nous suggérons un outil facile d'utilisation pour planifier une étude de fiabilité et pour utiliser le logiciel de planification lorsqu'il y a des contraintes. Nous présentons aussi un exemple pour illustrer l'analyse de données partir des plans proposés. *La revue canadienne de statistique* 39: 344–355; 2011 © 2011 Société statistique du Canada

## 1. INTRODUCTION

Reliability studies are widely used in medical and other contexts to assess both new and existing measurement systems. See Shoukri et al. (2004) for a review of the design issues, especially sample size, in such studies from a medical perspective. Burdick et al. (2005) in their book provide a more extensive review from an industrial perspective. Measurement systems, especially in the medical context, often include a large number of raters or operators who are one source of the variability seen when the same subject is measured more than once by different raters. In this article, we address the planning and analysis of efficient reliability studies when we can assume the raters used in the study are a random sample from a large population of possible raters. We adopt the

* *Author to whom correspondence may be addressed.*
*E-mail: shsteine@uwaterloo.ca*

following model for a single measurement on one subject that explicitly displays the sources of variation.

$$Y = P + O + M \tag{1}$$

The random variable $Y$ represents the measured value, $P$ is a random variable representing the true value of the characteristic, $O$ is a random variable representing the rater (operator) effect, and $M$ is a random variable, which represents the residual variation due to other sources in the measurement system. We make the usual normality and independence assumptions: $P \sim N(\mu, \sigma_p^2)$, $O \sim N(0, \sigma_o^2)$, $M \sim N(0, \sigma_m^2)$ and $P$, $O$, and $M$ are independent. In model (1), $\mu$ represents the overall mean measurement, $\sigma_p$ quantifies the variation due to differences in the actual characteristic values among subjects, $\sigma_o$ quantifies the variation due to differences among the raters, and $\sigma_m$ quantifies the variation due to other sources in the measurement system. We see the effects of this final source of variation when there are repeated measurements by a single rater on the same subject. The overall variation due to the measurement system is captured by $O + M$ with standard deviation $\sqrt{\sigma_o^2 + \sigma_m^2}$.

The intraclass correlation coefficient, denoted $\rho$, is a standard metric for assessing the quality of a measurement system (Donner and Eliasziw, 1987). The intraclass correlation is the correlation between two measurements on the same subject by different raters. From model (1), we have

$$\rho = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_o^2 + \sigma_m^2} \tag{2}$$

If $\rho$ is large the measurement system performs well since the variation due to real differences among the subjects is much larger than the variation introduced by the measurement system. Ideally $\rho > 0.8$ and any measurement system with $\rho < 0.5$ is unacceptable. In this article, in the assessment of a measurement system, we presume that $\rho$ is the primary parameter of interest. A parameter of secondary interest is $\delta = \sigma_m^2/(\sigma_o^2 + \sigma_m^2)$. If $\delta$ is close to 0, much of the overall measurement variation is due to rater to rater differences; if $\delta$ is close to 1, the rater to rater differences are relatively small compared to the other sources of variation in the measurement system.

The traditional plan for estimating $\rho$ is to select $k$ subjects and $r$ raters (at random) and then have each rater measure each subject once for a total of $N = kr$ measurements [1]. We denote such a plan by *SP*($k$,$r$). We extend model (1) to describe the data from this plan

$$Y_{ij} = P_i + O_j + M_{ij}, \quad i = 1, \dots, k; \quad j = 1, \dots, r \tag{3}$$

where we make the additional assumptions that there is independence both within and between $\{P_1, \dots, P_k\}$, $\{O_1, \dots, O_r\}$, and $\{M_{11}, \dots, M_{kr}\}$. In an industrial context, Burdick et al. (2005) consider the planning and analysis for *SP*($k$,$r$) using model (3).

Donner and Eliasziw (1987) and Walter et al. (1998) discuss optimal choices for $k$ and $r$ based on maximizing power at a given alternative for a test of the hypothesis $\rho = \rho_0$. The test is based on a one-way analysis of variance where in model (3), the rater effects are subsumed into the "other sources" terms, so the model becomes

$$Y_{ij} = P_i + M_{ij}^*, \quad i = 1, \dots, k; \quad j = 1, \dots, r \tag{4}$$

where we now assume that $\{P_1, \dots, P_k\}$ and $\{M_{11}^*, \dots, M_{kr}^*\}$ are independent. Walter et al. (1998) present an example where there are 30 subjects and 25 raters available and two further constraints. Due to respondent burden and time considerations, no subject can be measured more than four times and no rater can make more than six measurements. They recommend a plan in which there are five replicates of $SP(6,4)$ so that there are 20 raters, 30 subjects, and a total of 120 measurements made during the assessment.

Within each of the replicate standard plans, when model (3) is collapsed into model (4), the assumption of independence of the terms $\{M^*_{1j}, \ldots, M^*_{6j}\}$ for each rater $j$ is violated. If $\rho_0$ is relatively large, the correlation between any pair of these terms can be substantial. It is not clear whether the claimed power properties of the optimal design are met. In their discussion, Walter et al. (1998) acknowledge this issue and suggest that if there are substantial rater effects, more measurements are required to meet the power requirements.

In this article, we compare the precision of estimates of $\rho$ from various standard plans to the precision of the corresponding estimates from multiple replicates of a smaller standard plan. We use different (randomly selected) subjects and raters in each replicate of the *SP*. We denote this plan with $b$ replicates by $SP(k,r)^b$. With a replicated *SP*, we make a total of $N = krb$ measurements and use $kb$ different subjects and $rb$ different raters. The plan proposed by Walter et al. (1998), as described above, is $SP(6,4)^5$. In our comparisons, we select $k$, $r$, and $b$ so that the replicated plans have the same total number of measurements as the standard plans.

We start with a heuristic argument that suggests why replicate standard plans may be more efficient in estimating $\rho$ than are the corresponding standard plans. Next, we calculate the likelihood function, the Fisher information and a variance approximation for the MLE of $\rho$ for both standard and replicated standard plans. We then compare the two types of plans and show that multiples of $SP(2,2)$ have good properties. Next we provide a simple contour plot that can be used when planning a reliability study to determine the required number of multiples of $SP(2,2)$ to achieve a pre-specified precision for the estimate of $\rho$. We also give an example to demonstrate how to estimate $\rho$ and its standard error with available software. In the subsequent section, we look at cases when there are constraints on the design as described above. Finally, we provide a summary and discuss several issues that arise as a result of this work.

## 2. REPLICATED STANDARD PLANS

We motivate the good performance of replicated standard plans with the following argument. Since both the subjects and raters are random effects in model (3), we need a reasonable number of subjects and raters (Burdick et al., 2005) to estimate $\sigma_p^2$ and $\sigma_o^2$. In a two-way analysis of variance for $SP(k,r)$ (with no interaction), the three sums of squares for subjects, raters, and residual have $k-1$, $r-1$, and $(k-1)(r-1)$ degrees of freedom, respectively. These degrees of freedom roughly correspond to how well we can estimate the three variance components $\sigma_p^2$, $\sigma_o^2$, and $\sigma_m^2$. If $k$ and $r$ are large enough to produce good estimates of $\sigma_p^2$ and $\sigma_o^2$, then we have a very large number of degrees of freedom $(k-1)(r-1)$, to estimate $\sigma_m^2$. This lack of balance in the degrees of freedom results in inefficient estimation of $\rho$ especially when $\sigma_m^2$ is small. With a replicated plan, on the other hand, with a small number of subjects and raters for each replicate, the number of degrees of freedom within each replicate for each component of variance is similar. With $SP(2,2)^b$, the plan we recommend in the next section, there is one degree of freedom for each component of variance within each replicate. When we then combine information across the replicates using maximum likelihood estimation, there is a much better balance in the information about each component of variance than with a single large standard plan. As we show later, this results in more precise estimation of $\rho$ when $\rho > 0.5$, that is, for any reasonable measurement system.

It is not clear how to combine the among- and within-replicate sum of squares in order to use ANOVA for the replicated plans. Instead, we use likelihood methods for both planning and analysis. The likelihood for a replicated *SP* is the product of the likelihoods for each of the component *SP*s because of the independence assumptions and since each replicate uses different raters and different subjects. We start with likelihood for $SP(k,r)$ using model (3). We give the details of the derivation and the subsequent calculation of the Fisher information and variance approximations in the Appendix. We stack the data in a single column vector $y$ with the first $r$ elements corresponding to subject 1, the next $r$ to subject 2, and so on. Using model (3), the distribution of the corresponding random variable, $Y$, is multivariate normal with mean $\mu 1_{kr}$ where $1_{kr}$ is a vector of 1's of length $kr$ and variance–covariance matrix

$$\Sigma_{kr} = I_{kr}\sigma_m^2 + \begin{bmatrix} J_r\sigma_p^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & J_r\sigma_p^2 \end{bmatrix} + \begin{bmatrix} \begin{bmatrix} \sigma_o^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_o^2 \end{bmatrix} & \cdots & \begin{bmatrix} \sigma_o^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_o^2 \end{bmatrix} \\ \vdots & \ddots & \vdots \\ \begin{bmatrix} \sigma_o^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_o^2 \end{bmatrix} & \cdots & \begin{bmatrix} \sigma_o^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_o^2 \end{bmatrix} \end{bmatrix} \tag{5}$$

where $J_r$ is an $r \times r$ matrix of 1's.

Because the variance–covariance matrix has this simple form, the closed form expressions for the inverse and determinant of $\Sigma_{kr}$ given in the Appendix allows us to write the log-likelihood function as

$$\begin{aligned} l(\mu, \sigma_p^2, \sigma_o^2, \sigma_m^2\,|y) &= -\tfrac{1}{2}\ln(|\Sigma_{kr}|) - \tfrac{1}{2}(y-\mu)^T \sum_{kr}^{-1}(y-\mu) \\ &= -\tfrac{1}{2}\ln\left[(r\sigma_p^2 + k\sigma_o^2 + \sigma_m^2)(\sigma_m^2)^{kr-k-r+1}(r\sigma_p^2 + \sigma_m^2)^{k-1}(k\sigma_o^2 + \sigma_m^2)^{r-1}\right] \\ &\quad -\tfrac{1}{2}\left[b_1 \sum_{i,j}^{k,r}(y_{ij}-\mu)^2 + b_2 \sum_{i=1}^{k}\left(\sum_{j=1}^{r}(y_{ij}-\mu)\right)^2 + b_3 \sum_{j=1}^{r}\left(\sum_{i=1}^{k}(y_{ij}-\mu)\right)^2 + b_4\left(\sum_{i,j}^{k,r}(y_{ij}-\mu)\right)^2\right] \end{aligned} \tag{6}$$

where the constants $b_i$, $i = 1, \ldots ,4$ are functions of $k$, $r$, $\sigma_p$, $\sigma_o$, and $\sigma_m$ and are also given in the Appendix.

The log-likelihood of a replicated *SP* is the sum of the log-likelihoods from each of the individual component standard plans. Given the data from a *SP* or a replicated *SP*, we maximize the likelihood to estimate the model parameters.

We use the Fisher (expected) information to find approximations for the variance of the estimators of the three variance components $\sigma_m^2$, $\sigma_o^2$, and $\sigma_p^2$. From the log-likelihood given by (6), the corresponding Fisher information matrix is given by the negative of the expected value of the partial second derivatives with respect to $\mu$, $\sigma_m^2$, $\sigma_o^2$, and $\sigma_p^2$. The information about $\mu$ is separate from the information about $\sigma_m^2$, $\sigma_o^2$, and $\sigma_p^2$, that is, the information matrix is block diagonal. We found the long and messy expressions using Maple (Maplesoft, 2009). To avoid transcription errors, we used the Matlab (The Mathworks Inc., 2008) facility to accept pasted expressions from Maple (Maplesoft, 2009). In the Appendix, we give the expected values of the sums of squares in (6) that involve the data.

Denoting the lower $3 \times 3$ matrix of Fisher information matrix as $F$, the approximation for the variance–covariance matrix of the MLEs of the parameters $(\sigma_m^2, \sigma_o^2, \sigma_p^2)$ is given by $F^{-1}$. To find an approximation for the variance of the MLE of $\rho$, we use the delta method. The variance of the estimator of $\rho$ is approximately $D^T F^{-1} D$, where $D$ is the gradient vector $[\partial\rho/\partial\sigma_m^2,\ \partial\rho/\partial\sigma_o^2,\ \partial\rho/\partial\sigma_p^2]$. We provide Matlab (The Mathworks Inc., 2008) code (www.bisrg.uwaterloo.ca) for finding this variance approximation as a function of $\sigma_m^2$, $\sigma_o^2$, and $\sigma_p^2$.

## 3. COMPARISON AND RESULTS

In this section we compare *SP* and replicated *SP* plans with the same value of $N$, the total number of measurements, using the asymptotic standard deviation for the MLE of $\rho$ where we assume

$k \geq 2$, $r \geq 2$ to ensure we can estimate all the model parameters. In setting this constraint, we exclude the plans $SP(k,1)^b$ and $SP(1,r)^b$ that correspond to one-way designs.

We rank various plans by comparing the asymptotic standard deviation of the ML estimators. We start by reporting part of the results of an extensive simulation where we compared the approximations of the standard errors based on the Fisher information to those produced by simulation. We also examined the bias of the estimators which are asymptotically unbiased when both the number of raters and subjects goes to infinity. For example, we compared simulated and asymptotic results for $SP(2,2)^{15}$ and $SP(10,6)$ over a range of values for the underlying parameters. Both of these plans require a total of 60 measurements and $SP(2,2)^{15}$ uses 30 subjects and 30 raters. Without loss of generality, we set $\sigma_m^2 + \sigma_o^2 + \sigma_p^2 = 1$. For each simulation, we used 10,000 trials.

In Table 1, we see, for the cases reported, that the information-based and simulated standard errors for the MLEs are very close. In all cases, the MLEs have a negative bias that is smaller and negligible for the replicated plan. This bias is a well-known problem with maximum likelihood estimation based on model (4) (Swallow and Monahan, 1984).

More generally, for a wide variety of standard and replicated plans with different $k$ and $r$, the information-based approximations and simulated standard errors are very close as long as the total number of subjects and raters is reasonably large. In this case, for the replicated plans, the biases are negligible. We conclude that we can use the information-based approximations of the standard error as a basis for comparing various plans. For the remainder of the article, we refer to this approximation as the standard error.

To find good plans for specified values of $N$, we looked at all possible standard and replicated plans. For a grid of values for $\rho$ and $\delta$, we rank plans based on the standard error for the MLE of $\rho$. Not surprisingly, the best plan depends on the (unknown) parameter values. However, we found that the replicated $SP(2,2)^b$ plan is either the best or close to the best plan when we constrain $\rho$ to be $>0.5$, that is, when the measurement system is not very poor. We also found that unless $\delta$ is close to 1 (i.e., when $\sigma_o^2$ is small relative to $\sigma_m^2$), there is a replicated $SP$ that is much better than the best $SP$ for any particular choice of $\rho$ and $\delta$. We summarize the results in Figures 1 and 2 by

TABLE 1: Simulated biases and standard errors and approximated standard errors.

| | $SP(2,2)^{15}$ | | | | | $SP(10,6)$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\sigma_m^2$ | $\sigma_o^2$ | $\sigma_p^2$ | $\rho$ | | $\sigma_m^2$ | $\sigma_o^2$ | $\sigma_p^2$ | $\rho$ |
| True | 0.45 | 0.05 | 0.5 | 0.5 | True | 0.45 | 0.05 | 0.5 | 0.5 |
| Sim. mean | 0.426 | 0.074 | 0.475 | 0.474 | Sim. mean | 0.448 | 0.048 | 0.455 | 0.450 |
| Sim. SE | 0.138 | 0.090 | 0.200 | 0.145 | Sim. SE | 0.092 | 0.052 | 0.251 | 0.146 |
| Approx. SE | 0.161 | 0.123 | 0.204 | 0.137 | Approx. SE | 0.095 | 0.061 | 0.261 | 0.143 |
| True | 0.125 | 0.125 | 0.75 | 0.75 | True | 0.125 | 0.125 | 0.75 | 0.75 |
| Sim. mean | 0.125 | 0.125 | 0.726 | 0.733 | Sim. mean | 0.125 | 0.117 | 0.689 | 0.713 |
| Sim. SE | 0.045 | 0.072 | 0.224 | 0.091 | Sim. SE | 0.026 | 0.080 | 0.333 | 0.119 |
| Approx. SE | 0.046 | 0.071 | 0.227 | 0.083 | Approx. SE | 0.026 | 0.087 | 0.352 | 0.112 |
| True | 0.01 | 0.09 | 0.9 | 0.9 | True | 0.01 | 0.09 | 0.9 | 0.9 |
| Sim. mean | 0.010 | 0.089 | 0.868 | 0.891 | Sim. mean | 0.010 | 0.086 | 0.823 | 0.879 |
| Sim. SE | 0.004 | 0.035 | 0.241 | 0.046 | Sim. SE | 0.002 | 0.054 | 0.390 | 0.079 |
| Approx. SE | 0.004 | 0.035 | 0.245 | 0.041 | Approx. SE | 0.002 | 0.057 | 0.409 | 0.067 |

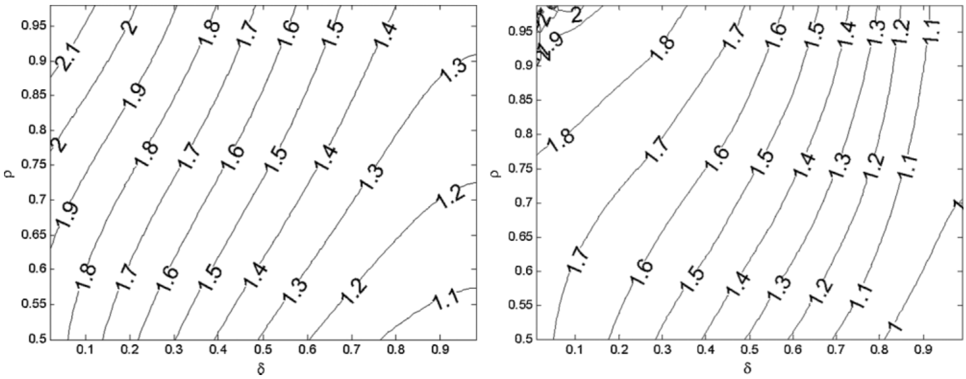

FIGURE 1: Efficiency of $SP(2,2)^{30}$ versus $SP(20,6)$ (left) and the best $SP$ (right), $N = 120$.

comparing the efficiency of the $SP(2,2)^b$ ($b = 15$, 30) plan to a specific standard $SP$ or the best $SP$ for the given the values of $\rho$ and $\delta$. We define efficiency as the ratio of standard errors (not variances, as is often done). The replicated $SP$ is preferred if the efficiency is $>1$.

We see in Figures 1 and 2 that replicated plan $SP(2,2)^b$ is substantially better at estimating $\rho$ than is any standard plan. When $\delta$ is small and $\rho$ is large the gains are substantial. The standard and replicated plans have similar efficiency when $\rho$ is small and $\delta$ is large. We find similar results for other values of $b$.

To determine how many replicates of the $SP(2,2)$ plan are required to give a desired standard error for the MLE of $\rho$, note that the Fisher information matrix for $SP(2,2)^b$ is $b$ times the Fisher information for $SP(2,2)$. And so the standard error of the MLE for $SP(2,2)^b$ is the standard error for $SP(2,2)$ divided by $\sqrt{b}$. In Figure 3, we plot contours of the standard error for $SP(2,2)$ as a function of $\rho$ and $\delta$. Note that for this plan, the standard error is sensitive to changes in $\rho$ but not $\delta$.

We illustrate the use of Figure 3 with the following example. Suppose we think $\rho \approx 0.8$ and $\delta \approx 0.5$. We want to estimate $\rho$ with a standard error of about 0.05. From Figure 3, the $SE(\hat{\rho})$ for $SP(2,2)$ is 0.27 when $\rho = 0.8$ and $\delta = 0.5$. Then, solving $0.27/\sqrt{b} = 0.05$ gives $b = 29$. The replicated plan $SP(2,2)^{29}$ will provide the desired precision. This plan has 58 raters, 58 subjects, and 116 measurements. We cannot safely apply the results in Figure 3 unless $b > 15$, that is, we have a least 60 measurements in total.

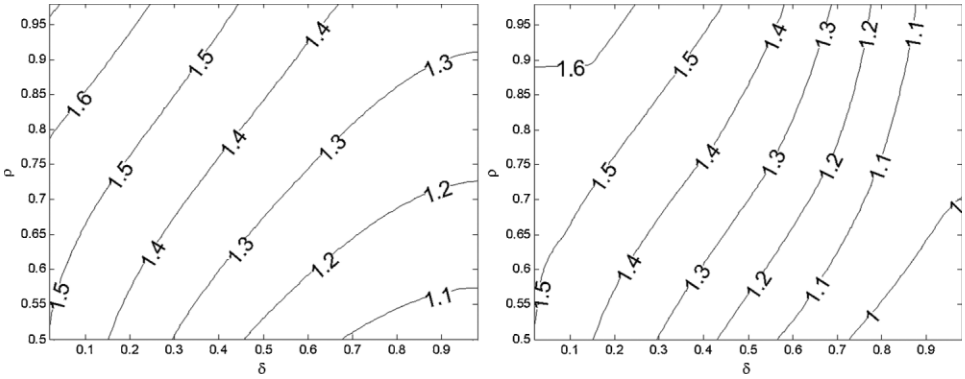

FIGURE 2: Efficiency of $SP(2,2)^{15}$ versus $SP(10,6)$ (left) and the best $SP$ (right), $N = 60$.

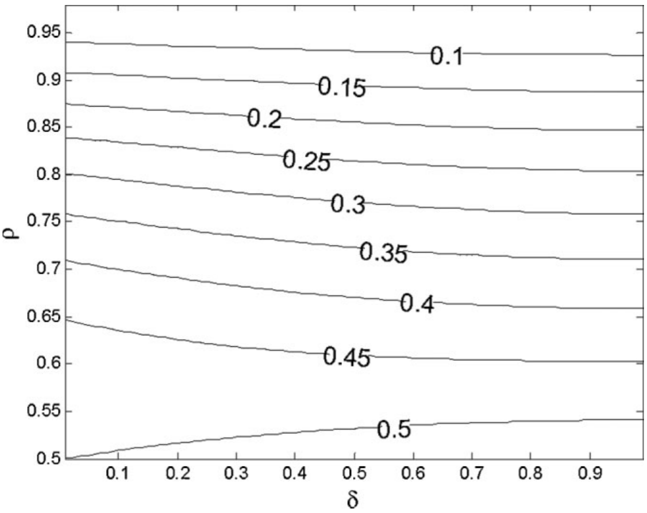

FIGURE 3: Contour plot of the standard error of $\hat{\rho}$ for $SP$(2,2).

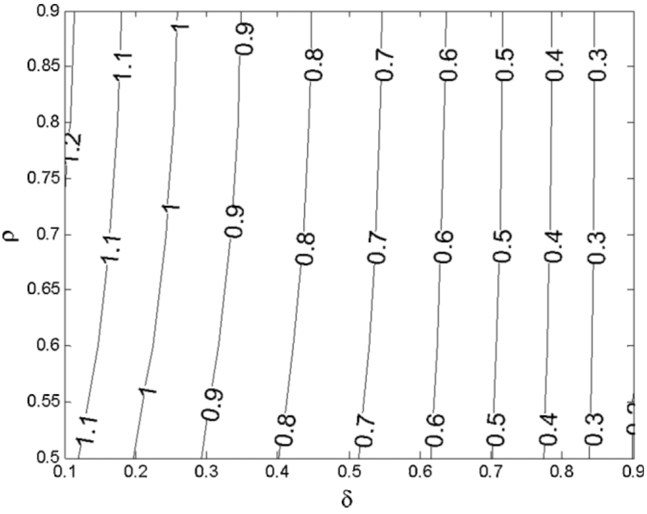

FIGURE 4: Contour plot of the standard error of $\hat{\delta}$ for $SP$(2,2).

We stated that the primary goal of the assessment study is to estimate $\rho$ and we recommend a replicated plan that is efficient in estimating the primary parameter. We can also look at $\delta = \sigma_m^2/(\sigma_m^2 + \sigma_o^2)$, a parameter that describes the relative contribution of the rater to rater differences to the variation due to the entire measurement system. In Figure 4, we show contours of the standard error of $\hat{\delta}$ for $SP$(2,2). We see that the standard error depends highly on $\delta$ and not on $\rho$. We also note, not surprisingly, that $SP$(2,2) provides far less information about $\delta$ than for $\rho$. For a plan with $b$ copies of $SP$(2,2), the standard error of $\hat{\delta}$ is $1/\sqrt{b}$ times the value given by Figure 4.

## 4. A NUMERICAL EXAMPLE

To illustrate the analysis of a replicated standard plan, we use data from the $SP(2, 2)^{10}$shown in Table 2. Note there are 20 subjects and 20 raters.

TABLE 2: Example data from a $SP(2,2)^{10}$ plan.

| | Replicate | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | | 2 | | 3 | | 4 | | 5 | | 6 | | 7 | | 8 | | 9 | | 10 | |
| Rater | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| First subject | 1.91 | 1.90 | 1.87 | 2.09 | 0.60 | 0.93 | 2.08 | 1.19 | 1.15 | 2.66 | 3.31 | 2.59 | 3.10 | 2.31 | 3.28 | 3.15 | -0.06 | -0.36 | 1.31 | 1.62 |
| Second subject | 2.22 | 2.28 | 0.11 | 0.08 | 2.98 | 2.65 | 4.79 | 3.92 | 2.34 | 2.96 | 0.68 | 0.14 | 4.32 | 3.64 | 1.94 | 1.13 | 2.76 | 2.60 | 0.01 | 0.17 |

TABLE 3: Best four plans satisfying the constraints.

| $k$ | $r$ | $n$ | $b$ | $SE(\sigma_m^2)$ | $SE(\sigma_o^2)$ | $SE(\sigma_p^2)$ | $SE(\delta)$ | $SE(\rho)$ |
|---|---|---|---|---|---|---|---|---|
| 6 | 4 | 1 | 5 | 0.0082 | 0.0213 | 0.2386 | 0.1166 | 0.0315 |
| 6 | 5 | 1 | 4 | 0.0079 | 0.0207 | 0.2654 | 0.1129 | 0.0332 |
| 3 | 2 | 2 | 10 | 0.0079 | 0.0260 | 0.2418 | 0.1380 | 0.0350 |
| 2 | 2 | 3 | 10 | 0.0075 | 0.0261 | 0.2951 | 0.1374 | 0.0389 |

The MLEs from this data are $(\hat{\mu}, \hat{\sigma}_m^2, \hat{\sigma}_o^2, \hat{\sigma}_p^2) = (1.958, 0.047, 0.132, 1.479)$. Thus, we have $\hat{\rho} = 0.892$. By substituting the estimates into the Fisher information, we get an approximate standard error for $\hat{\rho}$ of 0.0512. Since $\rho$ must lie between 0 and 1, it is best to derive the confidence interval on a transformed scale. We have found that the Fisher $z$-transform seems to work well. We let $\theta = 1/2\log((1+\rho)/(1-\rho))$ then $\partial\theta/\partial\rho = 1/(1-\rho^2)$ and $SE(\hat{\theta}) = (SE(\hat{\rho}))/(\partial\rho/\partial\theta) = (SE(\hat{\rho}))/(1-\hat{\rho}^2)$. So, for instance, to derive an approximate 95% confidence interval for $\rho$, we translate to the $\theta$ scale and find the approximate 95% confidence interval on this scale using $\hat{\theta} \pm 2SE(\hat{\theta})$. Then, we translate back to the $\rho$ scale. In the example, using the transformation, we have $\hat{\theta} = 1.43$ and $SE(\hat{\theta}) = 0.2505$. Thus, an approximate 95% confidence interval for $\theta$ is $1.43 \pm 0.5$. Translating back to the $\rho$ scale, we get the approximate 95% confidence interval (0.73, 0.96). In this case the confidence interval is wide because only 40 measurements were made. Matlab (The Mathworks Inc., 2008) code to find the MLEs and the approximate standard errors is available at (www.bisrg.uwaterloo.ca).

## 5. REPLICATED PLANS WITH CONSTRAINTS

In some contexts, there may be constraints on the number of raters or subjects available for the assessment. We let the maximum values be $R$ and $K$. Because of limited subject tolerance or because it is not possible to subdivide the test material into more than a few specimens, there may be a maximum number $A$ of measurements that can be made on any one subject (Walter et al., 1998). Similarly there may be a constraint on the number of subjects $B$ any one rater can assess. The total number of measurements $N$ satisfies the constraints $N \le RB$ and $N \le KA$. The replicated $SP$ plans proposed in this article are well suited to adapt to these sorts of constraints. A $SP(k, r)^b$ plan uses $kb$ subjects, $rb$ raters and each subject is assessed only $r$ times. By selecting $k$, $r$, and $b$ appropriately, we can find all replicated plans that satisfy the given constraints.

Matlab (The Mathworks Inc., 2008) code for comparing the efficiency of all replicated $SP$s satisfying the constraints is available on the web site www.bisrg.uwaterloo.ca. The inputs to the program are the constraints $N$, $K$, $R$, $A$, and $B$ as well as initial guesses for $\rho$ and $\delta$. If there is no limit to one of these design parameters, we set its value sufficiently high so that it has no impact. The program identifies all plans that satisfy the constraint including those with $n > 1$. Recall that $n$ is the number of times each rater measures each subject. The feasible plans are ranked using the approximate standard error of the MLE of $\rho$.

Walter et al. (1998) give an example where there are at most $K = 30$ subjects, at most $R = 20$ raters and each rater can assess at most $B = 6$ subjects. We know that $N \le 120$. There are 18 plans that satisfy the constraints with $N = 120$. Table 3 lists the four plans with smallest standard errors for $\hat{\rho}$ when $\rho = 0.9$ and $\delta = 0.5$. The single standard plan $SP$(6,20) satisfying the constraints has $SE(\rho) = 0.0553$. Note that there are plans with $n > 1$ that are close to optimal.

We investigated $SP(6, 4)^5$ over the range $0.7 \le \rho \le 0.99$, $\quad 0.1 \le \delta \le 0.9$ and found that it was uniformly the best plan satisfying the constraints with the smallest standard error for $\rho$.

The $SP(6, 4)^5$ is the plan recommended by Walter et al. (1998). Note however that here we are selecting this plan based on its properties using the appropriate model (3), not the one-way ANOVA model (4). The analysis with the incorrect model can yield substantial bias in the estimation of $\rho$ when there are substantial rater to rater differences.

In the above example, $SP(6, 4)^5$ consists of replicates of $SP(6,4)$, which is the smallest standard plan with a feasible value of $b$ that meets the constraints and uses exactly $N = 120$ measurements. We conjecture that this is generally true but cannot provide a proof. It is also interesting to note, for example, if $R = K = 40$ and $B = 6$, the plan $SP(3, 3)^{13}$ using 117 measurements produces a slightly smaller approximate standard error for $\rho$ than does $SP(6, 4)^5$ when $\rho = 0.9$ and $\delta = 0.5$ We suspect that higher powers of smaller standard plans are generally better because of the heuristic degrees of freedom argument presented earlier.

## 6. DISCUSSION AND SUMMARY

In our search for good measurement assessment plans, we also considered other ways to augment the traditional plan. We tried plans with the structure $SP(k, r) \times SP(s, 1)^b$ where $k$ and $r$ are small. In the augmented part, $b$ different raters measure each different subject once. This type of augmentation works well in the fixed rater effects case (Stevens et al. 2010) but with random rater effects, this structure never produced the best plan. More generally, we also considered augmented plans that consisted of a small *SP* together with any number of copies of another *SP* that could have different number of subjects and raters. For *N* fixed and some combinations of $\rho$ and $\delta$, these augmented plans were somewhat better than the replicated *SP* we recommend in this article. However, due to the additional complexity of implementing the plan and the difficulty of scaling the plan to a different total number of measurements we did not look at these plans any further.

For simplicity of exposition, we refer to raters throughout this article. The results are also applicable where the measurement system is automated (i.e., there are no rater effects) but there is a population of measurement devices. With model (1), we assume there is no interaction between subject and rater. That is, the effect of a particular rater is the same regardless of the subject. If, in fact, there is interaction so the effect of a rater changes from subject to subject, this extra variation is subsumed by *M*. It is not clear how this extra source of variation might change the results we have presented.

In summary, to assess the reliability of a measurement system where we assume raters are random effects and we ignore possible subject by rater interaction, we recommend a replicated standard plan as an improvement over the traditional plan. The replicated plan consists of many copies of the simple plan where two subjects are measured once by each of two raters. We showed that the replicated *SP* was more efficient for estimating $\rho$, the intraclass correlation coefficient, when $\rho > 0.5$, that is, when the measurement system is reasonable. The replicated *SP* plan $SP(2, 2)^b$ has other advantages over the *SP* since each subject is measured only two times regardless of the value of $b$. If there are constraints on the number of subjects, raters, and the number of times each subject can be measured or each rater measure, then we recommend investigating all feasible plans of the form $SP(k, r)^b$.

## APPENDIX

We can write the variance–covariance matrix given in (5) compactly as

$$\Sigma_{kr} = I_{kr}\sigma_m^2 + (I_k \otimes J_r)\sigma_p^2 + (J_k \otimes I_r)\sigma_o^2$$

where $I_a$ is an $a \times a$ identity matrix, $J_a$ is an $a \times a$ matrix of ones, and $\otimes$ is the Kronecker product. By direct multiplication, the inverse is

$$\Sigma_{kr}^{-1} = I_{kr}b_1 + (I_k \otimes J_r)b_2 + (J_k \otimes I_r)b_3 + J_{kr}b_4$$

where the constants $b_i$, $i = 1, \ldots, 4$ are functions of $k$, $r$, $\sigma_p^2$, $\sigma_o^2$, and $\sigma_m^2$

$$b_1 = \frac{1}{\sigma_m^2}, \quad b_2 = -\frac{\sigma_p^2}{\sigma_m^2(r\sigma_p^2 + \sigma_m^2)}, \quad b_3 = -\frac{\sigma_o^2}{\sigma_m^2(k\sigma_o^2 + \sigma_m^2)}$$

$$b_4 = \frac{\sigma_p^2\sigma_o^2(2\sigma_m^2 + r\sigma_p^2 + k\sigma_o^2)}{r^2\sigma_p^4\sigma_m^4 + 2r\sigma_m^6\sigma_p^2 + k^2\sigma_o^4\sigma_m^4 + 2k\sigma_m^6\sigma_o^2 + 3kr\sigma_m^4\sigma_p^2\sigma_o^2 + kr^2\sigma_p^4\sigma_o^2\sigma_m^2 + k^2r\sigma_p^2\sigma_o^4\sigma_m^2 + \sigma_m^8}$$

The determinant of $\Sigma_{kr}$ is

$$|\Sigma_{kr}| = (r\sigma_p^2 + k\sigma_o^2 + \sigma_m^2)(\sigma_m^2)^{kr-k-r+1}(r\sigma_p^2 + \sigma_m^2)^{k-1}(k\sigma_o^2 + \sigma_m^2)^{r-1}$$

The log-likelihood function for a standard plan with $k$ subjects and $r$ raters is thus given by (6). The expected values of the sums of squares in the likelihood needed to determine the Fisher information are

$$E\left[\sum_{i,j}^{k,r}(y_{ij}-\mu)^2\right] = kr(\sigma_p^2 + \sigma_o^2 + \sigma_m^2), \quad E\left[\sum_{i=1}^{k}\left(\left(\sum_{j=1}^{r}(y_{ij}-\mu)\right)^2\right)\right] = kr(r\sigma_p^2 + \sigma_o^2 + \sigma_m^2)$$

$$E\left[\sum_{j=1}^{r}\left(\left(\sum_{i=1}^{k}(y_{ij}-\mu)\right)^2\right)\right] = kr(\sigma_p^2 + k\sigma_o^2 + \sigma_m^2),$$

$$E\left[\left(\sum_{i,j}^{k,r}(y_{ij}-\mu)\right)^2\right] = kr(r\sigma_p^2 + k\sigma_o^2 + \sigma_m^2)$$

## BIBLIOGRAPHY

R. K. Burdick, C. M. Borror & D. C. Montgomery (2005). "*Design and analysis of gauge R&R studies: Making decisions with confidence intervals in random and mixed effects models*." ASA-SIAM Series on Statistics and Applied Probability, Philadelphia, PA.

A. Donner & M. Eliasziw (1987). Sample size requirements for reliability studies. *Statistics in Medicine*, 6, 441–448.

Maplesoft (2009) *Maple 13*. Waterloo, Ontario, www.maplesoft.com.

The MathWorks, Inc. (2008). *Matlab 7.7.0*. Natick, Massachusetts, www.mathworks.com.

M. M. Shoukri, M. H. Asyali & A. Donner (2004). Sample size requirements for the design of reliability study: Review and new results. *Statistical Methods in Medical Research*, 13, 251–271.

N. Stevens, R. Browne, S. H. Steiner & R. J. MacKay (2010). Augmented measurement system assessment. *Journal of Quality Technology*, 42, 388–399.

W. H. Swallow & J. F. Monahan (1984). Monte Carlo comparison of ANOVA, MIVQUE, REML and ML estimators of variance components. *Technometrics*, 26, 47–57.

S. D. Walter, M. Eliasziw & A. Donner (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine*, 17, 101–110.

*Received 15 September 2010*
*Accepted 1 February 2011*