Research Article

Statistics in Medicine

Received 22 June 2009, Accepted 14 September 2009 Published online 29 October 2009 in Wiley Interscience

(www.interscience.wiley.com) DOI: 10.1002/sim.3766

Optimal two-stage reliability studies

Ryan Browne, Stefan H. Steiner*† and R. Jock MacKay

The intraclass correlation is often used to assess the reliability of a measurement system. There is a considerable literature devoted to optimizing the standard assessment plan in which a number of subjects are measured repeatedly. We propose a two-stage investigation, here called a leveraged plan (LP), where in Stage I, we measure a number of subjects once. Then in Stage II, we select a subset of subjects with extreme initial measurements for repeated measurement. For a fixed total number of measurements, we show that the optimal LP provides a more precise estimate of the intraclass correlation coefficient than does the optimal standard plan (SP). We provide a table for finding the optimal LP given the true intraclass correlation and a specified precision for the estimate. For a fixed total number of measurements *N*, a nearly optimal LP makes roughly *N*/2 measurements in Stage I and then selects roughly *N*/6 extreme subjects to re-measure thrice each in Stage II. We also compare optimal leveraged with optimal SPs when there is a limit on the number of times each subject can be re-measured. Copyright © 2009 John Wiley & Sons, Ltd.

Keywords: intraclass correlation; measurement variation; leverage

1. Introduction

In medicine, good measurement systems are needed to support illness diagnosis and sound decision making. Large measurement variability hinders the ability to provide timely and effective treatment to patients who can benefit while making it more likely that inappropriate treatment is provided.

We usually assess measurement system reliability with a simple investigation where k subjects are randomly selected and then repeatedly measured n times. We refer to this as the standard plan (SP). A common model to describe this plan [1] is

$$Y_{ij}=A_i+E_{ij}, \quad i=1,2,\dots,k \quad \text{and} \quad j=1,2,\dots,n \tag{1}$$

where A_i is a random effect that describes the distribution of the true characteristic value for subject i and E_{ij} represents the distribution of the measurement error for the jth measurement on the ith subject. We assume that A_i is normal with mean μ and standard deviation σ_a and that E_{ij} is normal with mean zero and standard deviation σ_m. We also assume within subject independence among A_i and all E_{ij} and between-subject independence for all random variables.

The measurement variability can have many different sources. The repeated measurements can be made by different observers at different times under differing environmental conditions. With the simple one-way model (1), we assume that the true value of the characteristic for any subject does not change as the measurements are repeated and that all of the measurement variabilities are captured by the single parameter σ_m. Donner and Eliasziw [1] provide further discussion on these points.

To quantify the reliability of the measurement system, we use the intraclass correlation coefficient, defined as

$$\rho=\sigma_a^2/(\sigma_a^2+\sigma_m^2)=\sigma_a^2/\sigma_t^2 \tag{2}$$

where $\sigma_t=\sqrt{\sigma_a^2+\sigma_m^2}$ is the standard deviation of the measured values Y_{ij}. The intraclass correlation coefficient is the correlation between any two measurements on the same subject. A good measurement system has ρ close to 1 since then the measurement variation is small relative to the subject-to-subject variation.

With the SP, we estimate ρ using the sample intraclass correlation [1]. This plan and the corresponding analysis based on a one-way analysis of variance are widely used. See, for example, Doria *et al.* [2] who describe assessing the reliability of the compatible MRI scoring system for the evaluation of haemophilic knees, and Walter *et al.* [3] who describe a study to assess measurement reliability of the gross motor function of children with Down's syndrome.

Business and Industrial Statistics Research Group (BISRG), Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, Canada N2L 3G1

*Correspondence to: Stefan H. Steiner, Business and Industrial Statistics Research Group (BISRG), Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, Canada N2L 3G1.

†E-mail: shsteine@uwaterloo.ca

Donner and Eliasziw [1] and Bonnet [4] provide sample size calculations for the SP. If the total number of measurements $N=nk$ is fixed, we can compare the properties of a variety of SPs with differing n and k. Optimal SPs when N is fixed are described by Walter *et al.* [3] when the object is to maximize power in a hypothesis test and by Giraudeau and Mary [5] when the goal is to minimize the standard error of the estimate of the intraclass correlation.

In this article, we propose a two-stage leveraged plan (LP) for assessing the reliability of a measurement system. A similar mathematical situation was considered by Curnow [6] where to quantify the repeatability of lactation yield he considered an LP as described in Section 2 with $n=1$. Browne *et al.* [7–9] look at the design and use of LPs in a variety of industrial contexts, where periodic assessments of key measurement systems are required. The standard practice in the industrial context is to use an assessment plan with 10 or fewer parts (corresponding to subjects) and 6–9 repeated measurements per part. We compare the optimal form of the LP with the optimal SP and show that the former is substantially more efficient. We also derive the total number of measurements for optimal LPs based on a pre-specified precision of the estimate for the intraclass correlation and an assumed value for ρ. As the optimal LP is relatively insensitive to the unknown true value of ρ, we recommend a generic near-optimal LP. We also consider the situation where the number of repeated measurements on any one subject n is limited by practical or ethical considerations.

2. Leveraged plan for assessing measurement reliability

The LP is conducted in two stages:

Stage I: Sample b subjects at random and measure the characteristic of interest once on each subject to obtain a baseline sample. We denote the observed values by $\{y_{10}, y_{20}, \ldots, y_{b0}\}$.

Stage II: From the baseline sample, select k subjects using the observed measured values. We denote the indices of the k selected subjects using the set R. These k subjects are then repeatedly measured n (more) times each to give the additional data $\{y_{ij}, i\in R \text{ and } j=1,\ldots,n\}$. The total number of measurements in the LP is $N=b+nk$.

In Stage II, we sample k subjects with initial measurements that are extremely relative to the baseline average. For example, in an LP with $k=2$, we may pick the subjects with the minimum and maximum initial measurement in the baseline, or, with $k=4$, we may select the four most extreme (relative to the baseline average) subjects.

Next, we show how to determine the maximum likelihood estimate (MLE) of ρ and the corresponding standard error for data from an LP. Following Browne *et al.* [7], for any subject randomly selected for Stage I, the joint distribution of the baseline measurement from Stage I and the n repeated measurements in Stage II is

$$\begin{pmatrix} Y_0 \\ Y_1 \\ \vdots \\ Y_n \end{pmatrix} \sim \mathrm{N}\left(\mu\begin{bmatrix}1\\1\\ \vdots \\ 1\end{bmatrix}, \sigma_t^2\begin{bmatrix}1 & \rho\ldots & \rho\\ \rho & 1\ldots & \\ \vdots & \ddots & \vdots \\ \rho & \ldots & 1\end{bmatrix}\right)$$

Thus, the conditional distribution of the repeated measurements $\{Y_1, \ldots Y_n\}$ on a single subject given the baseline measurement $Y_0=y_0$ is

$$\left(\begin{array}{c|c} Y_1 & \\ \vdots & Y_0=y_0 \\ Y_n & \end{array}\right) \sim \mathrm{N}\left(\mu+\rho(y_0-\mu)\begin{bmatrix}1\\ \vdots \\ 1\end{bmatrix}, \Sigma=\sigma_t^2\begin{bmatrix}1-\rho^2 & & \rho(1-\rho)\\ & \ddots & \\ \rho(1-\rho) & & 1-\rho^2\end{bmatrix}\right) \tag{3}$$

In (3) the covariance matrix Σ has a special form that allows us to obtain the following results [10]:

$$\Sigma^{-1}=\frac{1}{\sigma_t^2(1-\rho)(1+n\rho)}\begin{bmatrix}1+\rho(n-1) & & -\rho\\ & \ddots & \\ -\rho & & 1+\rho(n-1)\end{bmatrix} \quad \text{and} \quad |\Sigma|=\sigma_t^{2n}(1-\rho)^2(1+n\rho)$$

Using the properties of covariance matrix Σ, we can write the conditional likelihood for the repeated measurements on a single subject. The measurements for one subject are independent of the measurements for other subjects; thus, the conditional likelihood for k subjects, each with n measurements, is the product of the individual likelihoods. The conditional log-likelihood for the n repeated measurements on k subjects is

$$l_r(\mu,\sigma_t^2,\rho|\text{ all } y_{i0} \text{ with } i\in R) = -\frac{nk}{2}\log\sigma_t^2-\frac{nk}{2}\log(1-\rho)-\frac{k}{2}\log(1+n\rho)$$
$$-\frac{1}{2}\frac{1}{\sigma_t^2(1-\rho)(1+n\rho)}\left\{(1+n\rho)k(n-1)MSW+n\sum_{i\in R}[\bar{y}_{i.}-\mu-\rho(y_{i0}-\mu)]^2\right\}$$

where $MSW=\sum_{i\in R}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2/k(n-1)$ and the subject average $\bar{y}_{i.}=(\sum_{j=1}^{n}y_{ij})/n$ omits the baseline measurement y_{i0}. A key point in the derivation of this likelihood is that the conditional distribution of the repeated measurements, given the initial measurement, depends on the initial measurement and not on how the subject is selected for Stage II.

Since the b subjects in the baseline sample (Stage I) are selected at random from the population, the marginal log-likelihood of the baseline data is

$$l_0(\mu,\sigma_t^2)=-\frac{b}{2}\log\sigma_t^2-\frac{1}{2\sigma_t^2}[(b-1)s_b^2+b(\bar{y}_b-\mu)^2]$$

where $\bar{y}_b$ and s_b are the average and standard deviation of the baseline measurements. The overall log-likelihood for the LP is thus

$$l(\mu,\sigma_t^2,\rho)=l_0(\mu,\sigma_t^2)+l_r(\mu,\sigma_t^2,\rho|\text{ all } y_{i0} \text{ with } i\in R) \tag{4}$$

To get the MLEs of μ, σ_t^2 and ρ, we maximize (4) numerically.

We can estimate the standard errors of the MLEs by inverting the Fisher information matrix and replacing the parameters with their estimates. Using the likelihood (4), the Fisher information matrix, i.e. the negative of the expectation of the partial second derivatives of the log-likelihood function with respect to (μ,σ_t^2,ρ), is given by

$$J(\mu,\sigma_t^2,\rho)=\begin{pmatrix} \dfrac{(1-\rho)nk+b(n\rho+1)}{\sigma_t^2(n\rho+1)} & 0 & \dfrac{nE[SC]}{\sigma_t(n\rho+1)} \\ 0 & \dfrac{b+nk}{2\sigma_t^4} & -\dfrac{nk\rho(n+1)}{2\sigma_t^2(n\rho+1)(1-\rho)} \\ \dfrac{nE[SC]}{\sigma_t(n\rho+1)} & -\dfrac{nk\rho(n+1)}{2\sigma_t^2(n\rho+1)(1-\rho)} & E\left[-\dfrac{\partial^2}{\partial\rho^2}l(\mu,\sigma_t^2,\rho)\right] \end{pmatrix} \tag{5}$$

where

$$E\left[-\frac{\partial^2}{\partial\rho^2}l(\mu,\sigma_t^2,\rho)\right]=\frac{kn^2}{2(1+n\rho)^2}+\frac{kn\rho(n+1)}{(1+n\rho)(1-\rho)^2}-\frac{kn}{2(1-\rho)^2}+\frac{nE[SSC]}{(1-\rho)(1+n\rho)}$$

$$SC=\sum_{i\in R}\left[\frac{Y_{i0}-\mu}{\sigma_t}\right] \quad\text{and}\quad SSC=\sum_{i\in R}\left[\frac{Y_{i0}-\mu}{\sigma_t}\right]^2$$

We estimate $E[SC]$ and $E[SSC]$ using the corresponding sample quantities. Note that due to the leveraged selection of subjects in Stage II, $E[SC]$ and $E[SSC]$ are not zero and k, respectively, as they would be with the random selection of subjects for Stage II. For comparison purposes, we give the Fisher information matrix for the SP in the Appendix.

For given values of b, k, and n, the optimal LP, i.e. the one where the asymptotic standard deviation of the MLE of ρ is minimized, has $E[SC]=0$ and $E[SSC]$ large [7]. As such, for Stage II of the LP, we recommend choosing an equal number of subjects on each side of the baseline average with extreme initial measurements. It is more important to ensure that $E[SSC]$ is large than $E[SC]$ close to zero. Thus, plans that select unequal numbers of subjects with large and small initial values are also good.

Since the distribution of the estimator for ρ is skewed (especially when ρ is close to the boundaries zero and one), we derive approximate confidence intervals on a transformed scale. We use the Fisher z-transform and let

$$\theta=\frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right),\quad \frac{\partial\theta}{\partial\rho}=\frac{1}{1-\rho^2} \tag{6}$$

Then

$$\text{Var}(\hat{\theta})\approx\text{Var}(\hat{\rho})\left(\frac{\partial\theta}{\partial\rho}\right)^2_{\rho=\hat{\rho}} \tag{7}$$

To create a confidence interval for ρ, we first derive a confidence interval for θ and then invert the limits using the inverse of the transformation. We illustrate the calculations with the following simple example. Suppose we conduct an LP where we select

26 subjects for Stage I (the baseline) and measure the characteristic of interest once for each of these subjects, yielding the data shown in Table I.

Subjects corresponding to the eight most extreme observations, four large and four small (bolded in Table I), were selected for Stage II and measured a further three times each. The Stage II data are given in Table II.

Using these data, the maximum of the log-likelihood given by (4) is $(\hat{\mu}, \hat{\sigma}_t^2, \hat{\rho}) = (7.81, 1.38, 0.88)$. From the inverse of the Fisher information given by (5) the standard error for $\hat{\rho}$ is 0.041. Transforming using (6), we get $\hat{\theta} = 1.376$, with standard error from (7) equal to 0.186, hence, an approximate 95 per cent confidence interval for θ is (1.02, 1.74). Transforming the endpoints back to the ρ scale, the approximate 95 per cent confidence interval for ρ is (0.77, 0.94). R code (R statistical freeware available at http://www.r-project.org/) for calculating the MLEs and approximate confidence intervals is available upon request from the authors.

3. Choosing a leveraged plan

In this section, we derive optimal LPs in terms of the baseline sample size (b), the number of subjects selected for Stage II (k), and the number of repeated measurements (n) made on each of the Stage II subjects. We start by guessing a value for ρ and specifying the desired maximum standard error of the estimate. We define optimal as the plan that has the fewest total number of measurements $N = b + kn$ that achieves the desired precision.

Table III gives the optimal LP in terms of (b, k, n) for a range of true ρ values. These plans were found by enumerating all possible plans and increasing the total number of measurements until a feasible plan was found. We express the precision condition as $z = SE(\hat{\rho})/\sqrt{\rho(1-\rho)(1-\rho^2)}$ so that we can cover more of the relevant range for each guessed value of ρ. The optimal plans are more stable if we use this transformed scale so that interpolation between the given values is reasonable. As expected the total number of measurements required increases as the precision condition on the standard error of the estimate becomes stricter.

When $\rho > 0.5$, we see in Table III that the optimal plans use approximately $N/2$ subjects in the baseline. Any reasonable measurement system has $\rho > 0.5$ and often ρ is close to one. If the plan is selected based on the total resources (number of measurements) available, then we suggest the generic plan $(b, k, n) = (N - 3 * \text{floor}(N/6), \text{floor}(N/6), 3)$, where floor($x$) is the largest integer not greater than x. With this plan, the baseline consists of roughly $N/2$ subjects and uses half of the total resources of the study.

In many medical applications [3], the number of repeated measurements (n) we can make on each subject is constrained by practical or ethical considerations. Fortunately, in the most common situation where the measurement system is reasonably good, say $\rho > 0.6$, the optimal plans, as shown in Table III, require only three additional measurements in Stage II, i.e. $n = 3$. Note that an LP with $n = 3$ measures some subjects once and others four times each (counting the baseline measurement). When we make the restriction $n = 1$, the optimal LP is very close to the SP, where in Stage II we measure all subjects in the baseline once more. Table IV gives results for the case where we fix $n = 2$. Compared with the results in Table III, we see that the constraint has little effect on the total number of measurements required. The best plans with $n = 2$ are nearly optimal overall.

To illustrate the application of these tables, we adapt the example given in Walter *et al.* [3] on the assessment of an instrument for measuring motor function in children with Down's syndrome. Walter *et al.* presented their results in the context of hypothesis testing. We instead specify a required maximum standard error. Suppose it was thought that the intraclass correlation was around 0.8, and that the largest acceptable standard error for the estimate of ρ was 0.05. With these values, $z = 0.05/\sqrt{0.8 * 0.2 * (1 - 0.8^2)} = 0.208$ or approximately 0.21. From Table III, the optimal LP plan in this case is $(b, k, n) = (45, 14, 3)$ with 87 total measurements. That is, with the optimal LP, we have a baseline sample of 45 subjects and, in Stage II, we select the 14 subjects with the most extreme initial measurements to measure three additional times each. The corresponding optimal

Table I. Baseline data.

9.19	8.24	7.31	**6.56**	**5.97**	7.56	**10.33**	**9.85**	6.94	**6.32**	7.44	7.09	6.86
6.77	7.98	8.15	8.72	**9.44**	7.29	7.45	**9.49**	**6.72**	7.72	7.67	9.07	6.87

Table II. Stage II data.

		Repeated measurement		
Subject #	Baseline measurement	#1	#2	#3
5	5.97	6.72	5.58	6.34
10	6.32	6.55	7.5	6.43
4	6.56	5.64	5.94	6.75
22	6.72	6.35	6.46	6.97
18	9.44	9.62	9.75	8.67
21	9.49	10.09	9.66	9.73
8	9.85	9.84	10.02	9.57
7	10.33	10.19	9.71	9.51

Table III. Optimal leveraged plans (b, k, n).

	$z=SE(\hat{\rho})/\sqrt{\rho(1-\rho)(1-\rho^2)}$							
ρ	0.36	0.33	0.3	0.27	0.24	0.21	0.18	0.15
0.1	(15,3,11)	(15,4,10)	(18,4,12)	(26,5,11)	(29,6,12)	(43,8,11)	(56,11,11)	(77,16,11)
0.2	(16,4,6)	(19,4,7)	(20,6,6)	(27,6,7)	(38,8,6)	(45,11,6)	(60,15,6)	(88,18,7)
0.3	(16,4,5)	(19,6,4)	(21,6,5)	(27,7,5)	(37,8,5)	(45,11,5)	(60,15,5)	(88,21,5)
0.4	(18,4,4)	(20,5,4)	(23,6,4)	(26,8,4)	(36,9,4)	(45,12,4)	(62,16,4)	(88,23,4)
0.5	(16,4,4)	(20,6,3)	(21,6,4)	(27,7,4)	(33,9,4)	(45,11,4)	(61,15,4)	(85,22,4)
0.6	(16,5,3)	(19,6,3)	(23,7,3)	(27,9,3)	(34,11,3)	(45,14,3)	(64,18,3)	(88,27,3)
0.7	(16,5,3)	(18,6,3)	(22,7,3)	(29,8,3)	(34,11,3)	(44,14,3)	(60,19,3)	(89,26,3)
0.75	(16,5,3)	(18,6,3)	(22,7,3)	(29,8,3)	(34,11,3)	(44,14,3)	(60,19,3)	(90,26,3)
0.8	(16,5,3)	(18,6,3)	(22,7,3)	(26,9,3)	(34,11,3)	(45,14,3)	(61,19,3)	(88,27,3)
0.85	(16,5,3)	(18,6,3)	(23,7,3)	(27,9,3)	(34,11,3)	(45,14,3)	(62,19,3)	(89,27,3)
0.9	(16,5,3)	(18,6,3)	(23,7,3)	(27,9,3)	(35,11,3)	(46,14,3)	(60,20,3)	(88,28,3)
0.95	(16,5,3)	(19,6,3)	(23,7,3)	(28,9,3)	(36,11,3)	(45,15,3)	(62,20,3)	(88,29,3)
0.99	(16,5,3)	(19,6,3)	(23,11,2)	(28,9,3)	(34,12,3)	(46,15,3)	(61,21,3)	(88,30,3)

b is the baseline size, k equals the number of subjects selected for Stage II and n is the # of repeated measurements.

Table IV. Optimal leveraged plans $(b, k, 2)$.

	$z=SE(\hat{\rho})/\sqrt{\rho}(1-\rho)(1-\rho^2)$							
ρ	0.36	0.33	0.3	0.27	0.24	0.21	0.18	0.15
0.1	(48,18,2)	(55,22,2)	(68,26,2)	(83,32,2)	(104,41,2)	(138,52,2)	(187,71,2)	(266,103,2)
0.2	(28,12,2)	(33,14,2)	(42,16,2)	(50,20,2)	(64,25,2)	(84,32,2)	(114,43,2)	(166,61,2)
0.3	(23,9,2)	(27,11,2)	(34,12,2)	(42,15,2)	(50,20,2)	(67,25,2)	(90,34,2)	(129,49,2)
0.4	(20,8,2)	(25,9,2)	(29,11,2)	(35,14,2)	(45,17,2)	(59,22,2)	(79,30,2)	(115,42,2)
0.5	(18,8,2)	(22,9,2)	(27,10,2)	(34,12,2)	(41,16,2)	(55,20,2)	(74,27,2)	(104,40,2)
0.6	(18,7,2)	(22,8,2)	(25,10,2)	(31,12,2)	(40,15,2)	(51,20,2)	(69,27,2)	(100,38,2)
0.7	(17,7,2)	(21,8,2)	(24,10,2)	(30,12,2)	(38,15,2)	(49,20,2)	(66,27,2)	(96,38,2)
0.75	(17,7,2)	(21,8,2)	(24,10,2)	(30,12,2)	(38,15,2)	(48,20,2)	(66,27,2)	(94,39,2)
0.8	(17,7,2)	(20,8,2)	(24,10,2)	(30,12,2)	(36,16,2)	(48,20,2)	(64,28,2)	(94,39,2)
0.85	(17,7,2)	(19,9,2)	(24,10,2)	(28,13,2)	(36,16,2)	(47,21,2)	(64,28,2)	(93,40,2)
0.9	(17,7,2)	(19,9,2)	(24,10,2)	(28,13,2)	(37,16,2)	(47,21,2)	(63,29,2)	(92,41,2)
0.95	(15,8,2)	(19,9,2)	(23,11,2)	(29,13,2)	(35,17,2)	(46,22,2)	(62,30,2)	(90,43,2)
0.99	(15,8,2)	(19,9,2)	(23,11,2)	(27,14,2)	(36,17,2)	(45,23,2)	(62,31,2)	(90,44,2)

SP requires 52 subjects each measured twice for a total of 104 measurements. In this case the optimal LP requires 17 fewer measurements than the SP to achieve the required precision. If we decided to restrict the maximum number of times each subject can be measured to three (i.e. $n=2$ in the LP), the optimal SP is unchanged and the optimal LP, given in Table IV, becomes $(b, k, n)=(48, 20, 2)$ with 88 total measurements. R code to help planning for other situations is available upon request from the authors.

When implementing an LP, we need to take care. Ideally, just like for an SP, the Stage I subjects and observers should be sampled randomly from the population of interest. If this is not the case, we may not obtain reasonable estimates of μ and σ_t^2. In addition, the measurements in both stages of an LP should reflect the normal measurement conditions. Otherwise, we may underestimate the measurement variation. Ideally, in Stage II, the repeated measurements should be spread out over time to allow the major sources of measurement variability to act. Finally, if possible, the Stage II measurements should be conducted shortly after Stage I. This helps to ensure that the properties of the measurement system are stable during the assessment.

4. Comparison of the standard and leveraged plans

In this section we further compare the performance of the optimal and recommended LP and the optimal SP. We choose the optimal plans to have the fewest total number of measurements that achieve a specified standard error for the estimator of ρ, whereas the recommended LP is the generic (i.e. not dependent on the true value of ρ) plan with $(b, k, n)=(N-3*\text{floor}(N/6), \text{floor}(N/6), 3)$ discussed in Section 3.

In the left panel of Figure 1, we show contours of the ratio N_{LP}/N_{SP} of the total number of measurements for the optimal LP, N_{LP}, over the number of measurements for the optimal SP for a range of values of N_{SP} and true values for ρ. We focus on the cases where $\rho \geqslant 0.5$ as this excludes very poor measurement systems. The optimal SPs match the optimal plans given in

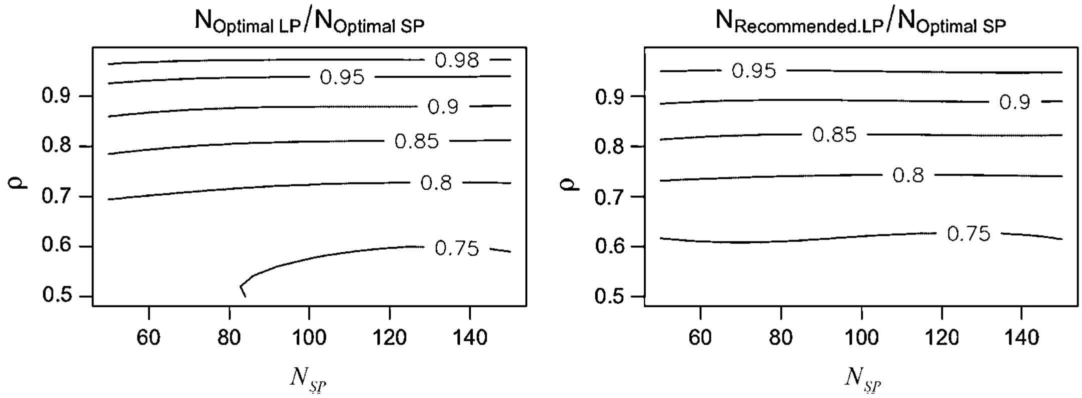

Figure 1. Comparison of LP Plans and the Optimal SP Plan, Contours give $N_{\mathrm{LP}}/N_{\mathrm{SP}}$.

Giraudeau and Mary [5]. To check that the asymptotic results are reasonable for the number of measurements used here, we also verified Figure 1 using simulation (not shown).

From the left panel of Figure 1, we see that the optimal LP requires fewer measurements than the optimal SP. For example, when ρ is near 0.7, the optimal LP requires only about 80 per cent of the total number of measurements needed by the optimal SP. The left panel of Figure 1 also suggests that the advantage of the optimal LP over the optimal SP increases slightly as the total number of measurements (N) increases. The right panel of Figure 1 is similar, but it compares the recommended (generic) LP with the optimal SP. We see a similar pattern, the recommended LP is substantial better than the optimal SP for most values of ρ. While here we compared the required total number of measurements to give a desired precision, the results also suggest that given the same total number of measurements, the optimal or recommended LP can provide a substantially more precise estimate of the intraclass correlation coefficient than the optimal SP.

As discussed in the previous section, we often limit n, the number of repeated measurements on each subject. A comparison of the optimal or recommended LP and SP with $n=2,3,4$ yields results essentially equivalent to those given in Figure 1. The only marked difference occurs for values of $\rho<0.5$ where the benefits of the LP over the SP are even more pronounced.

5. Discussion and conclusions

By adopting model (2) using either the standard or leveraged plans, we make a number of assumptions. In particular, we assume that both the measurement errors and the true characteristic values are normally distributed, the independence of the measurement errors and the true characteristic values and that the measurement variability does not depend on the subject (or the true value of the characteristic). We can check the assumption of normality of the measurement errors by examining the residuals (defined as the observed minus the subject average) from the repeated measurements in either the SP or Stage II of an LP. In both cases the number of degrees of freedom available is usually small. We can only indirectly check the assumption of normality for the true characteristic values by looking at the distribution of the baseline measurements. If both the measurement error and the true characteristics are normally distributed, the distribution of the observed measurements (from randomly selected subjects) should also be normal.

With either an SP or an LP, we repeatedly measure at least some of the subjects a number of times so that we have some power to detect whether σ_m, the measurement variability, is constant over the true characteristic values. To verify this assumption for an LP we would look for roughly equal variations in the residuals in the repeated measurements in Stage II stratified by the subject. In the SP plan we have repeated measurements on more subjects and a greater number of repeated measurements but in the LP plan we focus attention on extreme subjects where differences in the measurement variation are likely to be more pronounced. It is not clear which plan is better able to check the constant σ_m assumption.

A natural question is whether the LP is more sensitive to the model assumptions than the SP. This question is not easy to answer. One might expect problems because of the use of subjects with extreme baseline measurements. If the measurement error and the true values are not independent or the measurement variability is a function of the true characteristic value (a form of dependence), then with either plan, the meaning of ρ is not clear and we are unsure of what we are estimating. In a small factorial simulation study, Browne [11, Section 4.4] allowed the random variables A_i and E_{ij} to follow either a normal distribution or a t-distribution with five degrees of freedom (i.e. heavier tailed than the normal). Looking at all four combinations showed that the LP is not overly sensitive to the distributional assumptions and certainly no more so than the SP.

One additional risk in adopting the LP is the possibility of selecting extreme subjects for Stage II that are not representative of the measurement process. In practice, if a wild outlier is observed in the baseline measurements, we do not recommend the use of the corresponding subject in Stage II. Such an outlier may be due to either the process or the measurement system. We advise a separate study of this subject, because, if the extreme value is due to the measurement system, finding such an outlier in a small baseline study suggests that there may be a larger problem with the measurement system.

In some circumstances, such as with the assessment of a measurement system in current use, it may be reasonable to assume that Stage I of the leveraged reliability assessment plan has already been conducted. In fact, the number of subjects with a single measurement may be large. In this case, assuming that it is possible to select some extreme subjects from the previously measured ones, the LP is even more efficient. This situation is discussed in an industrial context by Browne *et al.* [8].

In this article, we considered the situation where each measurement is conducted by a different assessor. In this context, it is not possible to assess the measurement variation attributable to assessor-to-assessor differences. However, the leveraging approach can also be adapted to situations where the interest lies in partitioning the total variation into three components, say reflecting variation in the true dimensions, measurement variation due to repeated measurement and variation due to differences among the assessors [9].

In summary, we have proposed a new plan for a measurement reliability assessment plan that is more efficient than the previously described optimal plans [3, 5]. The LP is conducted in two stages. In the first stage, we measure a number of subjects once. Subjects in this baseline sample with extreme measurement are then selected to be repeatedly measured in the second stage of the LP.

The advantage of the LP over the SP depends on the true value of ρ and can be as large as a 25 per cent reduction in the required total number of measurements. The advantage of the LP over the SP extends also to the situation where there is a restriction on the number of repeated measurements per subject. Planning for an LP is simple since a nearly optimal LP with N total measurements is given by using roughly $N/2$ measurements to establish the baseline and then selecting roughly $N/6$ extreme subjects from the baseline to re-measure $n=3$ times each in Stage II.

Appendix

For the standard measurement reliability assessment plan, the Fisher information is

$$J_{\mathrm{SP}}(\mu,\sigma_t^2,\rho)=\begin{pmatrix} \dfrac{nk}{\sigma_t^2(n\rho+1-\rho)} & 0 & 0 \\ 0 & \dfrac{nk}{2\sigma_t^4} & -\dfrac{nk\rho(n-1)}{2\sigma_t^2(n\rho+1-\rho)(1-\rho)} \\ 0 & -\dfrac{nk\rho(n-1)}{2\sigma_t^2(n\rho+1-\rho)(1-\rho)} & \dfrac{nk(n\rho^2-\rho^2+1)(n-1)}{2(n\rho+1-\rho)(1-\rho)^2} \end{pmatrix}$$

We can estimate the standard error of the maximum likelihood estimator for the SP by inverting the Fisher information matrix, $J_{\mathrm{SP}}(\mu,\sigma_t^2,\rho)$, and replacing the parameters with their MLEs.

References

1. Donner A, Eliasziw M. Sample size requirements for reliability studies. *Statistics in Medicine* 1987; **6**:441–448. DOI: 10.1002/sim.4780060404.
2. Doria AS, Babyn PS, Lundin B, Kilcoyne RF, Miller S, Rivard GE, Moineddin R, Petterson H. Reliability and construct validity of the compatible MRI scoring system for evaluation of haemophilic knees and ankles of haemophilic children. Expert MRI Working Group of the International Prophylaxis Study Group. *Haemophilia* 2006; **12**:503–513.
3. Walter SD, Elaisziw M, Donner A. Sample size and optimal designs for reliability studies. *Statistics in Medicine* 1998; **17**:101–110. DOI: 10.1002/(SICI)1097–0258(19980115)17:1<101::AID-SIM727>3.0.CO;2-E.
4. Bonett DG. Sample size requirement for estimating intraclass correlations with desired precision. *Statistics in Medicine* 2002; **21**:1331–1335. DOI: 10.1002/sim.1108.
5. Giraudeau B, Mary JY. Planning a reproducibility study: how many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient. *Statistics in Medicine* 2001; **20**:3205–3214. DOI: 10.1002/sim.935.
6. Curnow RN. The estimation of repeatability and heritability from records subject to culling. *Biometrics* 1961; **17**:553–566.
7. Browne R, Steiner SH, MacKay RJ. Two stage leveraged measurement system assessment. *Technometrics* 2009; **51**:239–249.
8. Browne R, Steiner SH, MacKay RJ. Improved measurement assessment for processes with 100 per cent inspection. *Journal of Quality Technology* 2009; **41**:376–388.
9. Browne R, MacKay RJ, Steiner SH. Leveraged Gauge R&R Studies. *Technometrics* 2009; under review.
10. Dillon W, Goldstein M. *Multivariate Analysis Methods and Applications*. Wiley: New York, 1984.
11. Browne R. Leveraged plans for measurement system assessment. *Ph.D. Thesis*, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ont., Canada, 2009.