STATISTICS IN MEDICINE, VOL. 17, 101–110 (1998)

SAMPLE SIZE AND OPTIMAL DESIGNS FOR RELIABILITY STUDIES

S. D. WALTER,[1]* M. ELIASZIW[2,3] AND A. DONNER[3]

[1] *Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada L8N 3Z5*
[2] *The John P. Robarts Research Institute, London, Ontario, Canada N6A 5K8*
[3] *Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada N6A 5C1*

SUMMARY

A method is developed to calculate the required number of subjects k in a reliability study, where reliability is measured using the intraclass correlation ρ. The method is based on a functional approximation to earlier exact results. The approximation is shown to have excellent agreement with the exact results and one can use it easily without intensive numerical computation. Optimal design configurations are also discussed; for reliability values of about 40 per cent or higher, use of two or three observations per subject will minimize the total number of observations required. © 1998 John Wiley & Sons, Ltd.

Statist. Med., **17**, 101–110 (1998)

1. INTRODUCTION

Reliability studies are often used to evaluate the measurement properties of human observers or a test method. For instance, one may wish to assess the consistency of different clinicians in rating the functional status of their patients. In particular, we consider here studies where the intraclass correlation is used to measure reliability.

Following the notation of Donner and Eliasziw,[1] we suppose that n observations are made on each of k subjects, and that the jth observation Y_{ij} for subject i ($i = 1, 2, \ldots, k; j = 1, 2, \ldots, n$) is

$$Y_{ij} = \mu + a_i + \varepsilon_{ij}$$

where the random subject effects a_i are normally distributed with mean 0 and variance σ_A^2, or $N(0, \sigma_A^2)$, the measurement error ε_{ij} are $N(0, \sigma_\varepsilon^2)$, and the a_i and ε_{ij} terms are independent. We assume the subjects are randomly sampled from some population of interest. The intraclass correlation is defined as $\rho = \sigma_A^2/(\sigma_A^2 + \sigma_\varepsilon^2)$.

We emphasize designs where the k subjects are each rated by the same n raters (for inter-rater reliability). The same approach can, however, also be adopted when a single subject is observed repeatedly on each of several occasions (test–retest reliability), or when replicates of different

* Correspondence to: Dr. S. D. Walter, Department of Clinical Epidemiology and Biostatistics, McMaster University, HSC-2C16, 1200, Main St West, Hamilton, Ontario, Canada L8N 3Z5

CCC 0277–6715/98/010101–10$17.50

Received December 1996
Revised April 1997

types of measurement are obtained at different times.[2] In each of these cases, ρ can be estimated from a suitable one-way ANOVA. (We consider the possibility of using a two-way ANOVA in our Discussion). Higher values of ρ indicate greater reliability, in the sense that the measurement error is small relative to the between-subject variability. Guidelines for acceptable values of ρ have been proposed.[3]

Donner and Eliasziw have provided contours of exact power for selected values of k and n. The power results were also used later[4] to identify optimal designs that minimize study costs. The power calculations are based on numerical integration of incomplete beta functions, and the optimality results involve empirical fitting of polynomial regressions to the power contours, in conjunction with Lagrange multipliers to identify the optima.

The number of replicates n is often limited because of practical or logistical constraints. For example, in the study that motivated this work, physiotherapists were required to evaluate the gross motor function of children with Down's syndrome, using video tapes. Geographical constraints, respondent burden and other practicalities of scheduling meant that it was not possible to arrange for more than three or four therapists to assess a given child. Had live observations been made, it would have been unreasonable to increase the numbers of observers beyond this point, because of likely intimidation of the subject, particularly since assessments are often done in the child's home. Also, repetitive observations would probably lead to fatigue and undesirable learning or aversion effects. Similar considerations apply in many reliability studies, and n is typically limited by a maximum tolerable respondent burden, such as in an interview or physical assessment of clinical status.

In this paper, we develop a simple approximation that allows the calculation of required sample size for the number of subjects k, when the number of replicates n is fixed. The approximation uses a single formula, and avoids the intensive numerical work needed with the exact method. Furthermore, it permits the investigator to explore design options for parameter values not included in the figures provided by Donner and Eliasziw, such as for various settings of power. We are also able to identify optimal designs, where the total number of observations is minimized.

Use of the approximating formula for k also avoids another practical difficulty associated with use of the earlier exact results. For many designs of practical interest, the power contours do not clearly distinguish relevant design parameters, particularly when n is small. For typical values of n in the range 2–5, the power contours are essentially flat for a wide range of values of k, so visual identification of an efficient design is difficult.

Numerical comparisons of the approximate results with the exact results of Donner and Eliasziw show excellent agreement, suggesting that the approximate method can be recommended for practical use in the design of reliability studies.

METHOD

We consider a test of the null hypothesis H_0: $\rho = \rho_0$, where ρ_0 is a specified value of ρ. In the context of assessing reliability, ρ_0 would typically not be 0, because zero reliability is of no practical interest and is implausible anyway; here we would take ρ_0 to be the minimally acceptable level of reliability. In other situations, for example with cluster sample data, there may be interest in testing $H_0 : \rho = 0$. The alternative hypothesis is $H_1 : \rho > \rho_0$. H_0 is tested using $MS_A/MS_\varepsilon = [1 + (n-1)r]/(1-r)$ from the ANOVA, where MS_A and MS_ε are the mean squares for subjects and error, respectively, and where $r = (MS_A - MS_\varepsilon)/[MS_A + (n-1)MS_\varepsilon]$ is the

sample estimator of ρ. The critical value for the test statistic is $CF_{\alpha};v_1,v_2$, where $C = 1 + [n\rho_0/(1-\rho_0)]$ and $F_{\alpha};v_1,v_2$ is the $100(1-\alpha)$ per cent point in the cumulative F-distribution with (v_1,v_2) degrees of freedom, and where

$$v_1 = k - 1 \tag{1}$$

and

$$v_2 = k(n-1). \tag{2}$$

As described by Donner and Eliasziw,[1] at $\rho = \rho_1$ (where ρ_1 is a specific underlying value of ρ under H_1), the test of H_0 has power

$$1 - \beta = \Pr\{F \geqslant C_0 F_{\alpha};v_1,v_2\} \tag{3}$$

where β is the type II error and $C_0 = (1 + n\theta_0)/(1 + n\theta)$, with $\theta_0 = \rho_0/(1-\rho_0)$ and $\theta = \rho_1/(1-\rho_1)$. To solve (3), we note that if F has an F-distribution on (v_1,v_2) degrees of freedom, then Fisher's transformation[5] $z = \frac{1}{2}\ln F$ is distributed approximately as $N(\mu_z,\sigma_z^2)$, where

$$\mu_z = \frac{1}{2}\left(\frac{1}{v_2} - \frac{1}{v_1}\right). \tag{4}$$

and

$$\sigma_z^2 = \frac{1}{2}\left(\frac{1}{v_1} + \frac{1}{v_2}\right). \tag{5}$$

Substituting from (1) and (2) into (4) and (5) gives

$$\mu_z = \frac{(2k - kn - 1)}{2k(k-1)(n-1)} \tag{6}$$

and

$$\sigma_z^2 = \frac{(kn-1)}{2k(k-1)(n-1)}. \tag{7}$$

Fisher's transformation has been suggested to be most accurate for large v_1 and v_2, which, from (1) and (2), here implies large values of k. As shown later, the approximation is also adequate for smaller values of v_1 and v_2 that correspond to designs of practical interest.

Let z^* be the critical value of z. Then

$$z^* = \frac{1}{2}\left(\frac{1}{v_2} - \frac{1}{v_1}\right) + U_\alpha\left[\frac{1}{2}\left(\frac{1}{v_1} + \frac{1}{v_2}\right)\right]^{1/2} \tag{8}$$

where U_α is the $100(1-\alpha)$ per cent point in the cumulative unit normal distribution. From (3)

$$1 - \beta = \Pr[F \geqslant C_0 \exp\{2z^*\}]. \tag{9}$$

Combining equations (6)–(9), we have

$$C_0 \exp\left[\frac{(2k - kn - 1)}{2k(k-1)(n-1)} + 2U_\alpha\left(\frac{(kn-1)}{2k(k-1)(n-1)}\right)^{1/2}\right]$$

$$= \exp\left[\frac{(2k - kn - 1)}{2k(k-1)(n-1)} - 2U_\beta\left(\frac{(kn-1)}{2k(k-1)(n-1)}\right)^{1/2}\right]$$

which after simplification gives

$$\frac{2(U_\alpha + U_\beta)^2 (kn - 1)}{k(k-1)(n-1)} = (-\ln C_0)^2. \tag{10}$$

If k is large, we can replace the factor $(kn - 1)/k$ by n in (10), which then gives

$$\frac{2(U_\alpha + U_\beta)^2 n}{(k-1)(n-1)} = (-\ln C_0)^2. \tag{11}$$

After rearrangement (11) can be solved for k as

$$k = 1 + \frac{2(U_\alpha + U_\beta)^2 n}{(\ln C_0)^2 (n-1)}. \tag{12}$$

If the values of α and β have been fixed, and the other parameter values (ρ, ρ_0 and n) have been selected, all the quantities on the right hand side of (12) are known, and the required number of subjects k can be calculated.

The case of $n = 2$ also deserves special attention, as occurs with test–retest data. Fisher[6] suggested that, for the transformed intraclass correlation $z = \frac{1}{2}\ln[(1 + r)/(1 - r)]$, the variance should be taken as $1/[k - 3/2]$ instead of the usual value $1/[k - 3]$ for interclass correlations with $n > 2$. If one substitutes $\sigma_z^2 = 1/[k - 3/2]$ into (5) and uses a derivation similar to before, one obtains a modified version of (12) with leading term 3/2 instead of 1; so use of Fisher's modified variance leads to an increase of 0·5 in the required value of k. Note also that substitution of $n = 2$ into (8) gives $\sigma_z^2 = 1/[k - \{k/(2k - 1)\}]$, or approximately $1/[k - 1/2]$ if k is large.

The accuracy of the approximate formula (12) was verified against the values used to generate the exact power curves.[1] The exact calculations were made by selecting particular integral values of n and then determining k (non-integral) to satisfy the power requirements, at selected values of ρ_1, ρ_0, α and β; we took $U_\alpha = 1{\cdot}6449$ and $U_\beta = 0{\cdot}8416$. We then used (12) to determine an approximate solution k_{approx}, and compared it to k_{exact} used in the previous calculations. We also considered the accuracy of (12) using the modified variance $1/[k - 3/2]$ when $n = 2$, corresponding to the classical form of Fisher's transformation.

RESULTS

Table I shows specific numerical checks on the approximation, indicating very close agreement of the approximate and exact results over a wide range of values of ρ_1, ρ_0 and n. More extensive evaluation of the approximation, based on the 647 combinations of ρ_1, ρ_0 and n used to generate the exact power curves in Donner and Eliasziw,[1] revealed that $k_{\text{approx}} - k_{\text{exact}}$ had a median of 0·16 and a range from $-0{\cdot}40$ to 2·08. Only 27 per cent of these differences were negative. We conclude that the approximation can be used with some confidence in the design of actual studies, and that its bias is typically small and conservative.

For the case $n = 2$, we considered all combinations of $\rho_0 = 0(0{\cdot}2)0{\cdot}6$ and $\rho_1 = 0{\cdot}2(0{\cdot}2)0{\cdot}8$ with $\rho_1 > \rho_0$. The median value of $k_{\text{approx}} - k_{\text{exact}} + 0{\cdot}5$ (incorporating Fisher's modification to the variance) was 0·22, and the range was 0·18 to 0·28. Thus the modified approximation appears to work well in this case, giving values of k_{approx} that are again conservatively large.

Table I. Selected comparisons of approximate and exact methods for computing sample size k ($\alpha = 0{\cdot}05$, $\beta = 0{\cdot}20$)

ρ_0	ρ_1	n	k_{approx}	k_{exact}	Difference
0·0	0·2	20	5·05	5·00	0·05
0·0	0·4	10	4·31	4·30	0·01
0·0	0·4	3	16·37	16·06	0·31
0·0	0·6	2	13·87	14·13	− 0·26
0·0	0·8	10	2·00	2·20	− 0·20
0·2	0·6	2	26·71	26·99	− 0·28
0·2	0·8	2	8·70	8·94	− 0.24
0·4	0·6	5	35·05	34·01	1·04
0·8	0·9	10	22·61	21·72	0·89

k_{exact} was obtained from the Donner and Eliasziw tables. k_{approx} was calculated from equation (12)

Table II shows required values of k for typical values of n, and according to the values of ρ and ρ_0. (The values for $n = 2$ incorporate the 0·5 modification.) For example, suppose it has been decided that three replicates per subject are possible, and $\rho_0 = 0{\cdot}4$, $\rho_1 = 0{\cdot}6$, $\alpha = 0{\cdot}05$ and beta $= 0{\cdot}2$; then we require $k = 51{\cdot}5$ subjects, or 52 after rounding up.

The results in Table II indicate that the required sample size k depends critically on the values of ρ_1 and ρ_0, and on their difference in particular. So, for instance, considerably more effort is required to distinguish ρ values that differ by 0·1 compared to those with a difference of 0·2. Note also that larger samples are required in association with relatively small values of ρ, for a given difference $\rho_1 - \rho_0$.

Table III builds on these results by showing the optimal values of n, so that the total number of observations kn is minimized. These figures were derived by empirical inspection of the values of kn derived from (12), over a range of values of n, until the optimum was found. (Note that these optima pertain only to the situation where $\alpha = 0{\cdot}05$ and $\beta = 0{\cdot}20$.)

Table III indicates that smaller numbers of observations per subject are required if ρ_1 and ρ_0 are relatively large. In particular, for the range of reliability values likely to be of interest for many clinical measurement methods (say 0·5 or higher), the optimal value of n is between 2 and 5. Furthermore, if the minimally acceptable level of ρ (that is, the value ρ_0 defining H_0) is at least 0·4, then at most three measurements per subject are required, which often is achievable within the practical constraints of the study. Only if reliability is low and $\rho_1 - \rho_0$ is small does the optimal design specify a high value for n. If large numbers of replicates per subject are unattainable, one needs to adopt a sub-optimal design with a larger total number of observations.

Table IV shows the adequacy of the approximate optimal solutions, compared to designs suggested by the exact approach. The latter are drawn from Tables II–VI in Eliasziw and Donner[4] for situations where only the direct costs of the observations are taken into account (in their notation, where $R_2 = R_3 = 0$). The approximate results are obtained from (12), with k rounded up to the next integer value. The accuracy of the approximate method is again satisfactory for most practical purposes.

Table II. Estimates of sample size k, based on $\alpha = 0{\cdot}05$, $\beta = 0{\cdot}20$, for various values of ρ_0, ρ_1 and n

ρ_0	ρ_1								
	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9
$n = 2$									
0	615·6	151·9	70·0	35·9	22·0	14·4	9·7	6·6	4·4
0·1		591·2	142·8	60·6	32·2	19·1	12·0	7·7	4·8
0·2			543·7	128·2	53·0	27·2	15·5	9·2	5·3
0·3				476·2	109·0	43·5	21·4	11·4	6·1
0·4					393·1	86·6	32·9	15·1	7·1
0·5						300·3	62·6	22·0	8·8
0·6							205·4	39·1	11·7
0·7								117·1	18·4
0·8									45·8
$n = 3$									
0	225·1	60·2	28·1	16·4	10·7	7·4	5·3	3·8	2·7
0·1		251·8	64·8	29·2	16·4	10·2	6·8	4·6	3·0
0·2			261·1	64·8	28·1	15·1	9·0	5·6	3·4
0·3				251·8	60·2	25·1	12·8	7·1	4·0
0·4					225·1	51·5	20·3	9·6	4·7
0·5						183·9	39·6	14·4	5·9
0·6							133·1	26·1	8·0
0·7								79·7	12·8
0·8									32·5
$n = 4$									
0	122·9	35·3	17·5	10·8	7·4	5·4	4·0	3·1	2·3
0·1		156·7	42·4	20·0	11·7	7·6	5·3	3·7	2·6
0·2			177·8	45·9	20·6	11·5	7·1	4·6	2·9
0·3				183·3	45·2	19·4	10·2	5·9	3·4
0·4					172·4	40·4	16·4	8·0	4·1
0·5						146·6	32·3	12·0	5·1
0·6							109·7	21·9	6·9
0·7								67·5	11·1
0·8									28·3
$n = 5$									
0	80·2	24·5	12·8	8·2	5·8	4·4	3·4	2·7	2·1
0·1		114·5	32·2	15·7	9·5	6·4	4·5	3·3	2·3
0·2			139·4	37·0	17·1	9·7	6·2	4·1	2·7
0·3				150·8	38·0	16·6	9·0	5·3	3·1
0·4					147·0	35·1	14·4	7·2	3·8
0·5						128·4	28·7	10·8	4·7
0·6							98·1	19·9	6·4
0·7								61·5	10·3
0·8									26·1

Table II. (Continued)

ρ_0	ρ_1								
	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9
$n = 10$									
0	25·6	9·8	6·0	4·3	3·4	2·8	2·3	2·0	1·7
0·1		54·8	17·3	9·3	6·0	4·3	3·3	2·6	2·0
0·2			81·8	23·3	11·5	6·9	4·7	3·3	2·3
0·3				100·4	26·6	12·2	6·9	4·3	2·7
0·4					106·4	26·4	11·3	5·9	3·2
0·5						98·9	22·8	8·9	4·1
0·6							79·2	16·5	5·5
0·7								51·5	8·9
0·8									22·6
$n = 20$									
0	10·5	5·1	3·6	2·8	2·4	2·1	1·9	1·7	1·5
0·1		34·7	12·0	6·8	4·7	3·5	2·8	2·3	1·8
0·2			60·7	18·2	9·3	5·8	4·0	2·9	2·1
0·3				80·8	22·1	10·4	6·0	3·9	2·5
0·4					90·2	22·9	10·0	5·3	3·0
0·5						86·8	20·4	8·1	3·8
0·6							71·3	15·1	5·2
0·7								47·3	8·3
0·8									21·1

Table III. Optimal number of replicates n, to minimize total sample size kn ($\alpha = 0{\cdot}05$, $\beta = 0{\cdot}20$)

ρ_0	ρ_1								
	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9
0	26	13	8	6	5	4	3	3	3
0·1		8	6	5	4	4	3	3	3
0·2			5	4	4	3	3	3	3
0·3				4	3	3	3	3	3
0·4					3	3	3	3	3
0·5						3	3	3	2
0·6							3	3	2
0·7								2	2
0·8									2

EXAMPLE

As mentioned earlier, this work was motivated by discussions among investigators concerning the design of a reliability study for functional assessment by therapists of children with Down's syndrome. The instrument to be used, the GMFM (Gross Motor Functional Measure), had been previously developed and validated for use with children with cerebral palsy.[7] While development of motor function in children with Down's syndrome may be similar in some respects, there was concern that certain differences in physical and intellectual development might affect the reliability of the instrument in this context.

Table IV. Comparison of optimal values of k and n derived from the approximate and exact methods ($\alpha = 0{\cdot}05$, $\beta = 0{\cdot}20$)

ρ_0	ρ_1	k		n		kn	
		Exact	Approximate	Exact	Approximate	Exact	Approximate
0	0·2	8	8	12	13	96	104
	0·4	7	7	6	6	42	42
	0·6	7	6	3	4	21	24
	0·8	4	4	2	3	8	12
0·2	0·4	36	46	5	4	180	184
	0·6	12	16	4	3	48	48
	0·8	6	6	3	3	18	18
0·4	0·6	51	52	3	3	153	156
	0·8	9	10	3	3	27	30
0·6	0·8	26	27	3	3	78	81

Exact values are derived from Eliasziw and Donner,[4] restricted to integers. Approximate values are calculated from equation (12), and with k rounded up to the next integer value

Practical limitations meant that each child could only be evaluated by at most four therapists. The investigators were hoping for inter-rater reliability of at least 0·85, and had determined that reliability of 0·7 or higher would be acceptable. Hence we defined $H_0 : \rho_0 = 0{\cdot}7$ and $H_1 : \rho_1 = 0{\cdot}85$, with $\alpha = 0{\cdot}05$ and $\beta = 0{\cdot}2$.

The reliability work was done in the context of a population-based validation study which involved children from early intervention programmes for Down's syndrome in the province of Ontario. To increase generalizability, it was intended to include as many different therapists as possible. Preliminary enquiries suggested the availability of at least 15 therapists, and up to 25, who would be willing to take part. Each therapist was required to commit a total of one half-day of working time, which was sufficient to assess six children. Within the timeframe and other logistical constraints of the study, it was expected that approximately 30 subjects would be available.

Using (12) with the input parameters, the required number of subjects for various values of n, and the total number of observations kn were:

n	2	3	4
k	42·4	29·2	25·0
kn	83·8	87·6	100·0

Assuming 30 children could take part, 15 therapists observing 6 children each implies 3 observations per child; with 30 children and 20 therapists, 4 observations per child could be made. (Incidentally, these parameter values also permit some balance in the assignment of therapists to subjects.)

Although $n = 3$ is technically sufficient to meet the power requirement with 30 children (because the calculated $k = 29{\cdot}2$), it was decided to use $n = 4$ instead. This design allows some

'slack' for missing data or dropout of therapists. The design selected was not optimal in terms of minimizing kn, because $n = 2$ would be preferred with the selected values of ρ_1 and ρ_0; however, the optimal design would have required 43 children, which was more than were available in the relevant time frame.

DISCUSSION

The method presented here allows a relatively simple calculation of the required number of subjects k for a given number of observations per subject n, and other parameter values to specify the null and alternative values of reliability, significance level and power. Numerical comparison reveals that the approximation has sufficient agreement with the previous exact calculations to be of practical use. Note that we retained decimal values in our tabulated results on sample sizes and in the example, even though only integer values (after rounding up) would be required in practice. At the integer level, the agreement of our approximate results with the exact formulae is even better.

One of the limiting factors in executing reliability studies in many settings is the difficulty of arranging for the replicated observations. In clinical situations, for instance, there are typically only a few specialists available in one place (for example, a hospital or clinic) who are willing to participate in such a study and who are qualified to make the observations. Replication is also limited by the tolerance of the subjects. In reliability studies on laboratory methods, these problems are perhaps less acute, but even here there are limitations on the amount of laboratory resources (people and equipment) available, and it may not be possible to separate the experimental material (for example, tissue biopsies) into more than a few parts for replicated testing.

Accordingly, the design of reliability studies is often constrained to a limited range of potential values for n. In the Down's syndrome example, we indicated how the practical constraints were identified to define candidate values for n, and corresponding values of k. Similar exploration is indicated in other studies.

An interesting finding from the results of this paper is that the total number of observations is minimized with a relatively small value of n, as long as the true reliability is reasonably high (Table III). In many biomedical studies, reliability of at least 40 per cent is required to provide a clinically useful method of measurement, and values much higher than 40 per cent are often desirable. For such situations, one can uniformly recommend only two or three observations per subject. The optima in such cases are fortuitously in concordance with the practical limitations that often limit the value of n. Note, however, that these conclusions refer to optima at the conventional values $\alpha = 0{\cdot}05$ and $\beta = 0{\cdot}20$, and also only take the direct costs of observation into account. Modification may be needed in other cases, for instance when there is an additional cost associated with the recruitment of each subject.

As discussed previously by Donner and Eliasziw, the method used here is simplified by considering only between- and within-subject variation as components of reliability, with a corresponding one-way ANOVA for the data. In some circumstances, it might be possible to allow for more than one factor in the analysis, for instance in a design with full crossing of subjects and observers. Our approach to sample size calculation under a one-way ANOVA formulation also provides approximate guidelines to the two-way ANOVA situation, especially if rater effects are small. In general, a two-way analysis that incorporates rather effects into the definition of reliability will yield a required sample size that is a function of the rater effect. In actual data sets, if a true rater effect exists but is ignored, the error mean square is artifactually inflated, the

reliability coefficient is negatively biased, and the required sample size is increased. The more appropriate two-way analysis has correspondingly higher power, by removing between-rater variance from the error.

Kraemer[8,9] has considered reliability in a two-way ANOVA context, but only included subject and error effects in her definition of reliability, and excluded rater effects; thus her sample size expressions are independent of rater effects. A treatment of sample size calculation in the two-way ANOVA cases, taking rater effects explicitly into account, seems needed, but note that in practice it is often difficult to obtain prior estimates of the magnitude of the rater effect. For our example, other (unpublished) work has shown the GMFM measure to have inter-rater reliability of approximately 0·96. This high level of reliability justifies the use of the one-way ANOVA, in which the effects of specific raters are ignored. Lacking such information in other situations, the approximate method given in this paper for the one-way formulation seems to be a practical compromise.

ACKNOWLEDGEMENTS

We thank Dr. Peter Rosenbaum and Dianne Russell for helpful comments concerning the example in an early draft of this paper. Dr. Walter holds a National Health Scientist (Health Canada) award, and the work was partially supported by grants from the Natural Sciences and Engineering Research Council, Canada, to all three authors.

REFERENCES

1. Donner, A. P. and Eliasziw, M. 'Sample size requirements for reliability studies', *Statistics in Medicine*, **6**, 441–448 (1987).
2. Haggard, E. R. *Intraclass Correlation and the Analysis of Variance*, Dryden Press, New York, 1958.
3. Landis, J. R. and Koch, G. G. 'The measurement of observer agreement for categorical data', *Biometrics*, **33**, 159–174 (1977).
4. Eliasziw, M. and Donner, A. P. 'A cost-function approach to the design of reliability studies', *Statistics in Medicine*, **6**, 647–655 (1987).
5. Johnson, N. L. and Kotz, S. *Distributions in Statistics*: *Continuous Univariate Distributions 2*, Wiley New York, 1970, Chapter 26.
6. Fisher, R. A. *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh, 1925.
7. Russell, D., Rosenbaum, P., Cadman, D., Gowland, C., Hardy, S. and Jarvis, S. 'The Gross Motor Function Measure: a measure to evaluate the effects of physical therapy', *Developmental Medicine and Child Neurology*, **31**, 341–352 (1989).
8. Kraemer, H. C. 'The small sample non-null properties of Kendall's coefficient of concordance for normal populations', *Journal of the American Statistical Association*, **71**, 608–613 (1976).
9. Kraemer, H. C. and Korner, A. F. 'Statistical alternatives in assessing reliability, consistency and individual differences for quantitative measures: applications to behavioural measures of neonates', *Psychological Bulletin*, **83**, 914–921 (1976).

