STATISTICS IN MEDICINE, VOL. 6, 441–448 (1987)

SAMPLE SIZE REQUIREMENTS FOR RELIABILITY STUDIES

ALLAN DONNER AND MICHAEL ELIASZIW

Department of Epidemiology and Biostatistics, The University of Western Ontario, London, Ontario N6A 5B7, Canada

SUMMARY

This paper provides exact power contours to guide the planning of reliability studies, where the parameter of interest is the coefficient of intraclass correlation ρ derived from a one-way analysis of variance model. The contours display the required numbers of subjects k and number of repeated measurements n that provide 80 per cent power for testing $H_0: \rho \leqslant \rho_0$ versus $H_1: \rho > \rho_0$ at the 5 per cent level of significance for selected values of ρ_0. We discuss the design considerations of these results.

KEY WORDS Sample size Reliability Analysis of variance Intraclass correlation

INTRODUCTION

The estimation of reliability is a common feature of scientific experimentation since all measurements are subject to error, particularly those made by humans. As discussed by Shrout and Fleiss,[1] measurement error can seriously affect statistical analysis and interpretation; it therefore becomes important to assess the amount of such error by calculation of a reliability coefficient. A frequently adopted model to investigate reliability is

$$Y_{ij} = \mu + a_i + e_{ij}, \tag{1}$$

where μ is the grand mean of all measurements Y in the population, a_i reflects the effect of the characteristic under measure for subject i, and e_{ij} is the error of measurement, $j = 1, 2, \ldots, n$; $i = 1, 2, \ldots, k$. The error of measurement may result from both the measuring device itself and the conditions surrounding the measurement. We assume the term a_i remains constant across the repeated measurements on the same person.

Suppose we assume further that the person effects $\{a_i\}$ are normally and identically distributed with mean zero and variance σ_A^2, the errors $\{e_{ij}\}$ are normally and identically distributed with mean zero and variance σ_e^2, and the $\{a_i\}$, $\{e_{ij}\}$ are completely independent. Then the population intraclass correlation coefficient is $\rho = \sigma_A^2/(\sigma_A^2 + \sigma_e^2)$ and we use the one-way analysis of variance (ANOVA) as a framework for drawing inferences concerning ρ. The ANOVA corresponding to (1) appears in Table I, where the sample intraclass correlation

$$r = \frac{\text{MSA} - \text{MSW}}{[\text{MSA} + (n-1)\text{MSW}]} = \frac{F-1}{F+n-1}$$

estimates ρ. A large value of r implies greater variability among individuals than within individuals, i.e. that the repeated observations on a given subject show stability. The actual value of r estimates the proportion of total variance accounted for by subject to subject variation, and we may therefore

0277–6715/87/040441–08$05.00

Received May 1986
Revised November 1986

Table I. Analysis of variance for a reliability model

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Among subjects	$k-1$	SSA	$\text{MSA} = \text{SSA}/(k-1)$	MSA/MSW
Within subjects	$k(n-1)$	SSW	$\text{MSW} = \text{SSW}/[k(n-1)]$	

$$\text{SSA} = \sum_{i=1}^{k} n(\bar{Y}_{i.} - \bar{Y}_{..})^2 \qquad \text{SSW} = \sum_{i=1}^{k}\sum_{j=1}^{n} (Y_{ij} - \bar{Y}_{i.})^2$$

$$\bar{Y}_{i.} = \sum_{j=1}^{n} Y_{ij}/n \qquad \bar{Y}_{..} = \sum_{i=1}^{k}\sum_{j=1}^{n} Y_{ij}/kn$$

regard it as a measure of reliability. With dichotomous, rather than continuous data, r is equivalent to the Kuder–Richardson formula 20 reliability coefficient (Kraemer and Korner[2]).

Although intraclass correlation as a measure of reliability has garnered much attention in the literature (e.g. Shrout and Fleiss,[1] Fleiss *et al.*[3]), very little has appeared with regard to sample size requirements. In this paper we consider the exact values of k and n required to test H_0: $\rho = \rho_0$ versus H_1: $\rho > \rho_0$, where ρ_0 is a specified criterion value of ρ. Kraemer and Korner[2] have considered this problem for the case $n = 2$ (i.e., test–retest data), and Kraemer[4] has considered it for the case in which either k or n is large. These authors' results, however, are approximate rather than exact and are based on a two-way rather than a one-way ANOVA model. A further limitation is their provision of only the required value of k for fixed n; the results below deal with power requirements as both k and n vary.

We note the use of the estimator r in many contexts that involve repeated observations in each of several groups. Haggard[5] gives three such examples:

(i) a design with k subjects, each evaluated by n judges;
(ii) a design with a single subject evaluated n times on each of k occasions by the same judge;
(iii) a design in which a single subject provides k types of measurements, each type replicated at n different times by a single judge.

Case (i) consists of replication of the judges and r estimates interjudge concordance. In case (ii), r estimates the stability of the subject who provides a number of samples. In case (iii), r estimates the consistency of the single judge over intervals of time. In each case, interest focuses on one source of variation – judges, tests, trials – and r is an appropriate measure of reliability.

To simplify discussion, however, we consider r in this paper to represent the ratio of among subject variability to total variability. Thus, throughout we refer to k as the 'number of subjects' and n as the 'number of measurements per subject'; we assume it understood that our results apply to a broad range of investigations that involve data collected on k samples of n repeated observations, and with interest on a single source of variation.

METHOD

We assume interest focuses on the test H_0: $\rho \leqslant \rho_0$ versus H_1: $\rho > \rho_0$ at a chosen level of significance α and with power $1-\beta$. Selection of the criterion value ρ_0 depends on a choice of a minimum value of ρ that the investigators consider acceptable. The choice $\rho_0 = 0$ often is not

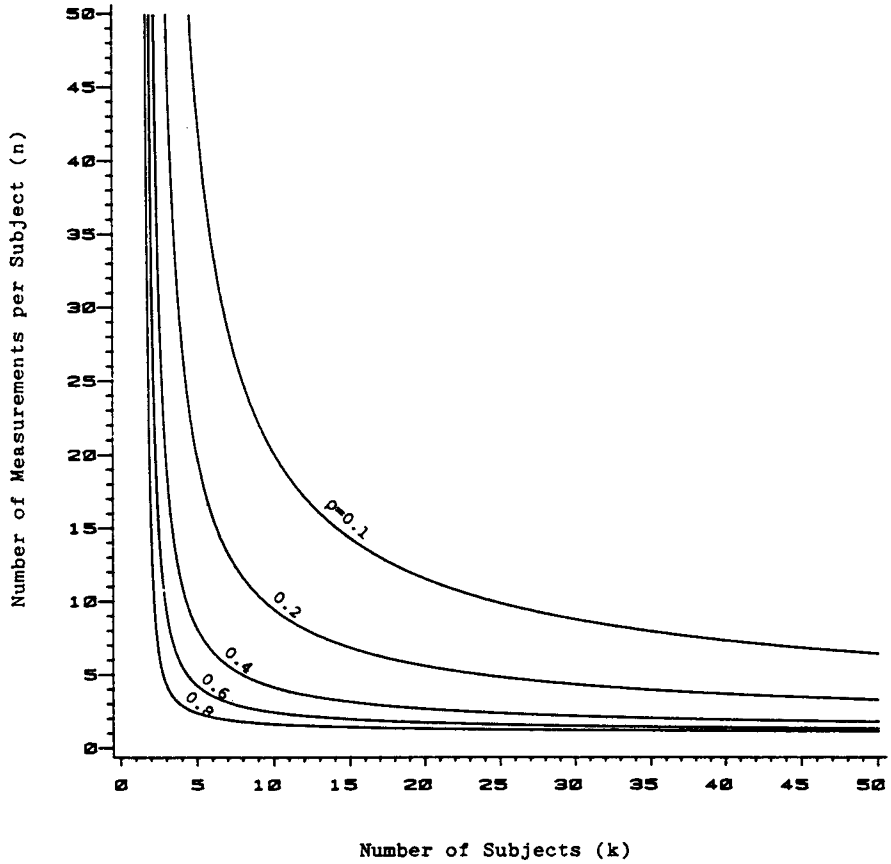

Figure 1. Power contours for testing H_0: $\rho = 0{\cdot}0$ versus $H_1: \rho > 0{\cdot}0$ at $\alpha = 0{\cdot}05$, $\beta = 0{\cdot}20$

directly relevant, since rejection of $H_0: \rho = 0$ simply provides reassurance at the chosen α level of greater variation among repeated observations on the same subject than among different subjects – a fact which we can usually take for granted. Landis and Koch[6] have characterized values of reliability coefficients as follows: slight (0–0·20), fair (0·21–0·40), moderate (0·41–0·60), substantial (0·61–0·80), and almost perfect (0·81–1·00). Although arbitrary, these divisions provide useful benchmarks. For example, an investigator who wishes to demonstrate a 'substantial' level of reliability would, according to these guidelines, test H_0: $\rho \leqslant 0{\cdot}60$ versus H_1: $\rho > 0{\cdot}60$.

One performs the test $H_0: \rho \leqslant \rho_0$ versus $H_1: \rho > \rho_0$ by reference of the value of F in Table I to the quantity $CF_\alpha; v_1, v_2$, where $C = 1 + [n\rho_0/(1-\rho_0)]$, and $F_\alpha; v_1, v_2$ is the tabular value of F with v_1, v_2 degrees of freedom at the α per cent level of significance. One may calculate the power of this test (Scheffe[7]) as

$$1 - \beta = Pr\{\, F \geqslant C_0\, F_\alpha; v_1, v_2 \,\} \tag{2}$$

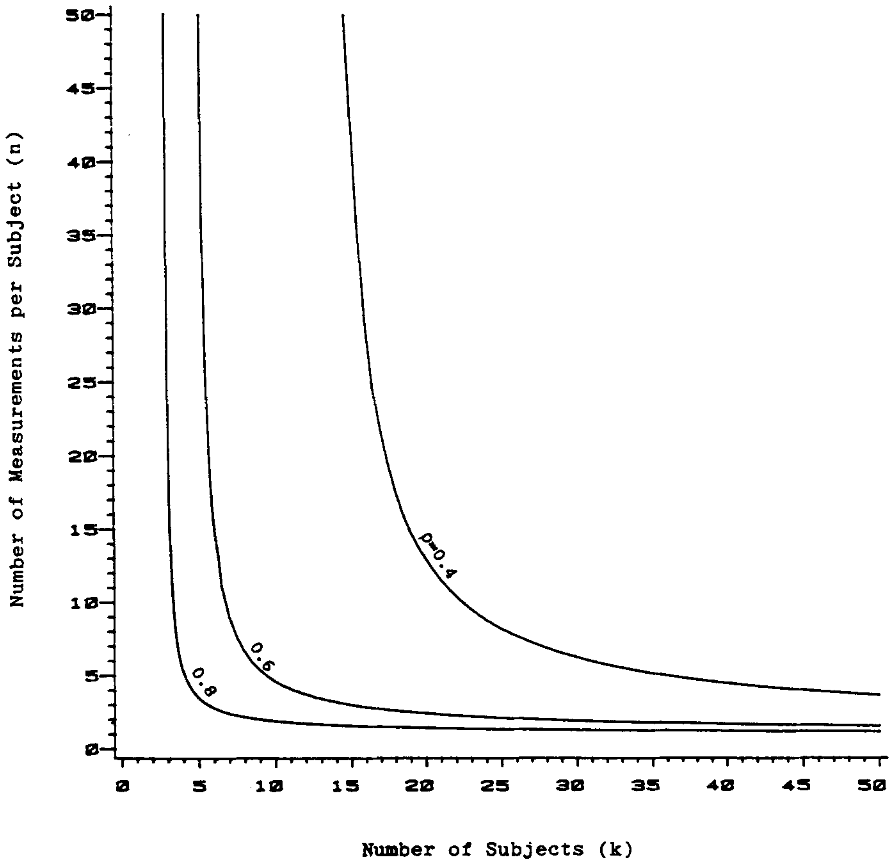

Figure 2. Power contours for testing H_0: $\rho = 0{\cdot}2$ versus H_1: $\rho > 0{\cdot}2$ at $\alpha = 0{\cdot}05$, $\beta = 0{\cdot}20$

where

$$\begin{aligned} v_1 &= k-1 \\ v_2 &= k(n-1) \\ C_0 &= (1+n\theta_0)/(1+n\theta) \\ \theta_0 &= \rho_0/(1-\rho_0) \\ \theta &= \rho/(1-\rho). \end{aligned}$$

Our investigation sought to generate contours of equal power to test $H_0: \rho = \rho_0$ versus $H_1: \rho > \rho_0$ in terms of k and n for fixed values of ρ_0 and ρ at $\alpha = 0{\cdot}05$. Since conventional power levels are set at 80 per cent or more, we fixed equation (2) at 0·80 and 0·90 for this purpose. We then generated the contours numerically for $\rho_0 = 0, 0{\cdot}2, 0{\cdot}4, 0{\cdot}6, 0{\cdot}8$ and selected values of $\rho > \rho_0$. For reasons of space we present the results only for $1-\beta = 0{\cdot}80$; upon request, we will make available corresponding results for $1-\beta = 0{\cdot}90$.

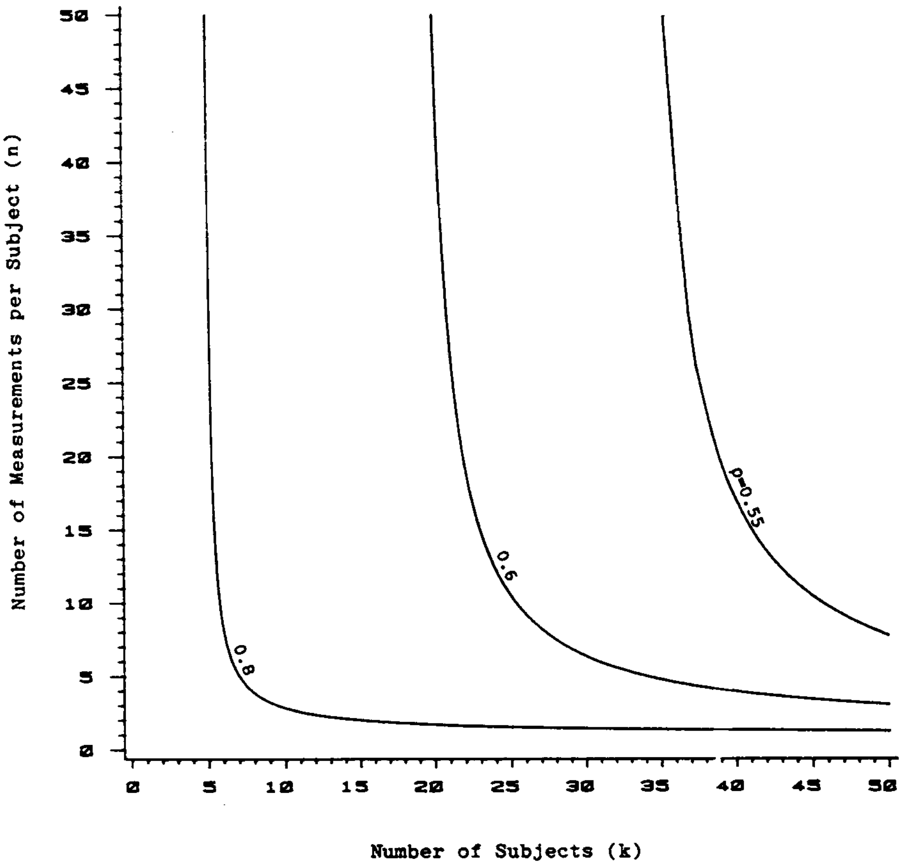

Figure 3. Power contours for testing H_0: $\rho = 0{\cdot}4$ versus H_1: $\rho > 0{\cdot}4$ at $\alpha = 0{\cdot}05$, $\beta = 0{\cdot}20$

COMPUTATIONS

All programs were written in Fortran using the RM/FORTRAN compiler, and were run on an IBM PC. For each chosen value of ρ, ρ_0 and $1 < k \leqslant 50$, and from equation (2) with use of an iterative method of successive bisection, we calculated to two decimal places the minimum value of n that satisfies the power requirements. We calculated the exact probabilities and percentage points of the F distribution with use of the computer algorithms for the incomplete beta distribution given in Kennedy and Gentle.[8]

RESULTS

The results appear in Figures 1–5 for $\rho_0 = 0, 0{\cdot}2, 0{\cdot}4, 0{\cdot}6$, and $0{\cdot}8$, respectively. Each figure shows the values of k and n that correspond to a power of 80 per cent for a test of H_0: $\rho = \rho_0$ versus H_1: $\rho > \rho_0$ at $\alpha = 0{\cdot}05$. For example, suppose we wish to demonstrate that $\rho > 0{\cdot}2$, i.e. in planning an investigation we characterize measurement reliability as at least 'fair'. If we wish 80 per cent certainty for achieving a significant result at the 5 per cent level when $\rho = 0{\cdot}4$, then Figure 2 shows that we require an n of about 13 when the number of available subjects k is 20. With 40 subjects

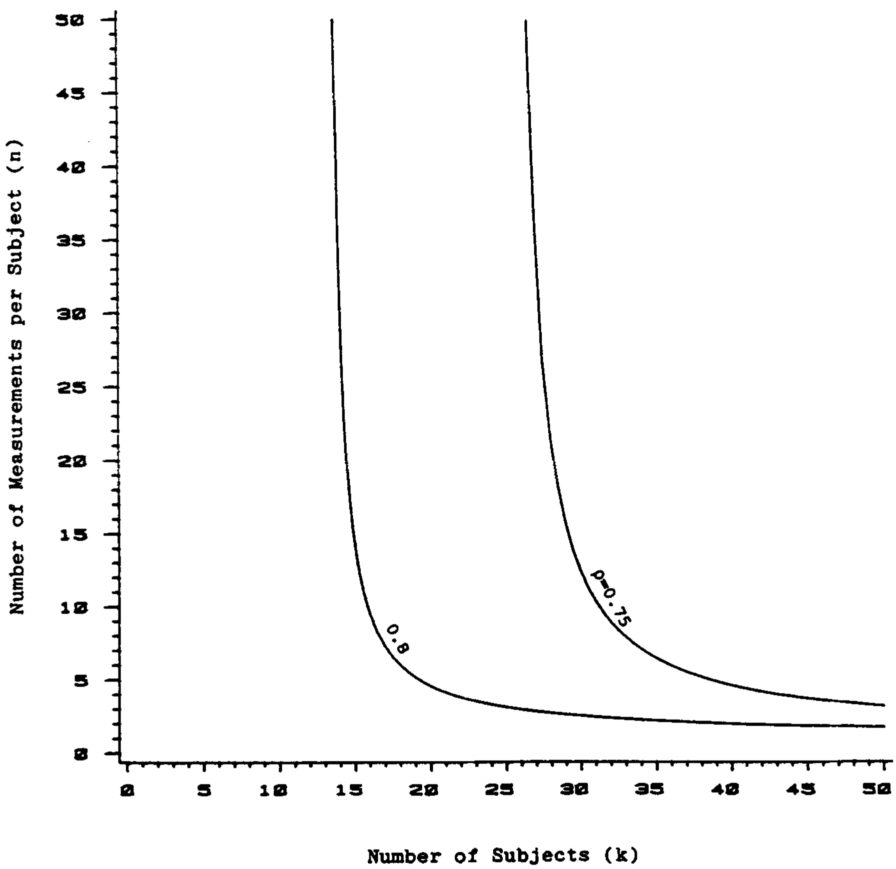

Figure 4. Power contours for testing H_0: $\rho = 0{\cdot}6$ versus H_1: $\rho > 0{\cdot}6$ at $\alpha = 0{\cdot}05$, $\beta = 0{\cdot}20$

available, $n = 5$ measurements per subject will suffice to achieve the desired power. Our actual choice of a (k, n) combination depends, of course, on the relative difficulty and cost of recruitment of subjects as compared to attainment of replicate measurements.

Figures 1–5 also reveal some interesting results concerning the relative influence of k and n on the achieved power. For example, we see the tendency towards a 'threshold' level of k beyond which any increase in k, with n held constant, brings very little return. Thus if $\rho_0 = 0{\cdot}4$, Figure 3 shows that at $\rho = 0{\cdot}8$, 15 subjects each with two measurements provide essentially the same power as 50 subjects each with two measurements. At $\rho = 0{\cdot}6$, 40 subjects with $n = 3$ provide about the same power as 50 subjects with $n = 3$. This feature of the contours reflects a general property of the estimator r: an increase in n for fixed k provides more information than an increase in k for fixed n.

The results also show that the required value of n for a given k increases very rapidly as k declines. At $\rho_0 = 0{\cdot}2$ and $\rho = 0{\cdot}6$, for example, a decrease in k below 10 results in a steep increase in the value of n required to maintain the power at 80 per cent. Thus one use of these results is as a guide in choice of the minimum number of subjects required to achieve fairly stable power to test H_0.

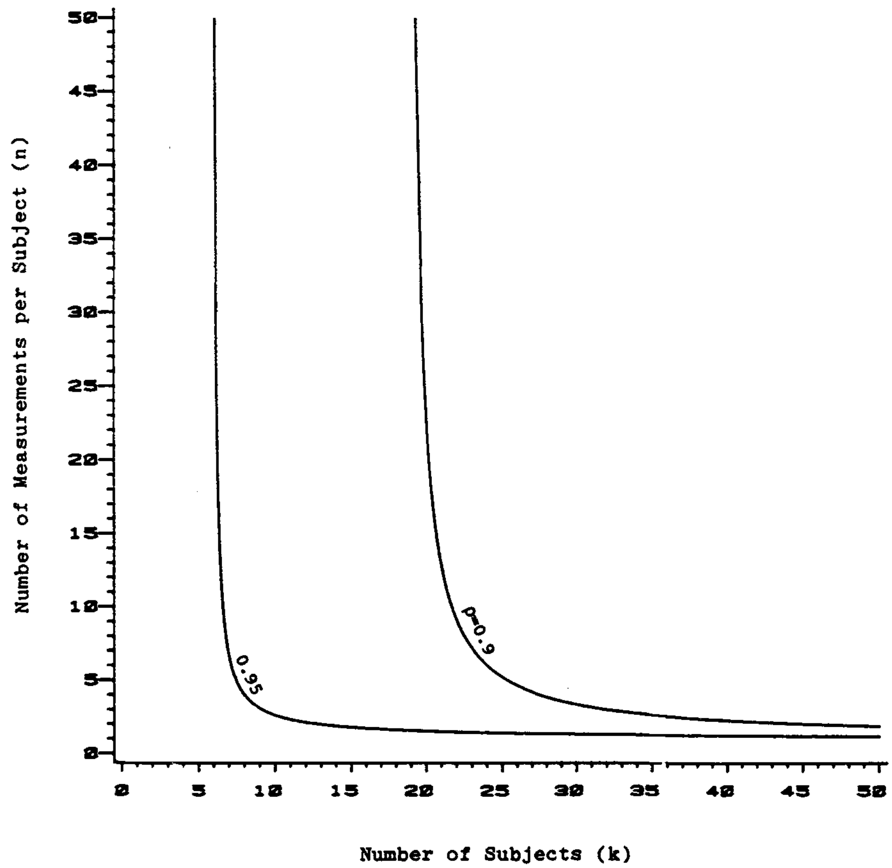

Figure 5. Power contours for testing $H_0: \rho = 0{\cdot}8$ versus $H_1: \rho > 0{\cdot}8$ at $\alpha = 0{\cdot}05$, $\beta = 0{\cdot}20$

DISCUSSION

The power contours presented here rely on an underlying one-way analysis of variance model. Such a model implies that systematic differences among the n measurements on a given subject are not separable from random error. If, for example, n different judges or examiners make the n measurements, the within subject sum of squares shown in Table I combines the effects due to judges and to random error. An alternative model for the reliability study is two-way analysis of variance, which partitions the within subject sum of squares into a between-judges and a residual component sums of squares. If a study is designed to estimate both the variation among judges as well as variation among subjects, then the results presented here do not apply and the two-way model is appropriate. The advantage of the one-way model is its simplicity, with some loss in precision compared with a two-way model that has large inter-judge differences. It follows that the sample size requirements of Figures 1–5 overestimate the true sample size requirements when one adopts a two-way model for the analysis, i.e. the results presented here are conservative. Shrout and Fleiss[1] provide further discussion of the factors influencing the choice of a reliability model.

It is useful in the analysis of a reliability study to construct confidence limits for the reliability coefficient. Let F_U denote the tabulated value of the F distribution with $k-1$ and $k(n-1)$ degrees of freedom that cuts off the proportion α in the upper tail. Then an exact one-sided $100(1-\alpha)$ per cent confidence interval for r is

$$r \geqslant (F - F_U)/[F + (n-1)F_U]$$

If this lower bound indicates acceptable reliability, then one may rely on single measurements with confidence. Fleiss[9] provides further discussion on the analysis aspects of reliability studies.

ACKNOWLEDGEMENTS

This research was supported in part by the National Health Research and Development Program through a National Health Fellowship to M. Eliasziw. Dr. Donner's research was supported by the Natural Science and Engineering Research Council of Canada.

REFERENCES

1. Shrout, P. E. and Fleiss, J. L. 'Intraclass correlations: uses in assessing rater reliability', *Psychological Bulletin*, **86**, 420–428 (1979).
2. Kraemer, H. C. and Korner, A. F. 'Statistical alternatives in assessing reliability, consistency and individual differences for quantitative measures: application to behavioural measures of neonates', *Psychological Bulletin*, **83**, 914–921 (1976).
3. Fleiss, J. L., Slakter, M. J., Fischman, S. L., Park, M. H. and Chitton, N. W. 'Inter-examiner reliability in caries trials', *Journal of Dental Research*, **58**, 604–609 (1979).
4. Kraemer, H. C. 'The small sample non-null properties of Kendall's coefficient of concordance for normal population', *Journal of the American Statistical Association*, **71**, 608–613 (1976).
5. Haggard, E. R. *Intraclass Correlation and the Analysis of Variance*, Dryden Press, New York, 1958.
6. Landis, J. R. and Koch, G. G. 'The measurement of observer agreement for categorical data', *Biometrics*, **33**, 159–174 (1977).
7. Scheffe, H. *The Analysis of Variance*, Wiley, New York, 1959.
8. Kennedy, W. J. and Gentle, J. E. *Statistical Computing*, Marcel Dekker, New York, 1980.
9. Fleiss, J. L. *The Design and Analysis of Clinical Experiments*, Wiley, New York, 1986.