HIGHER-MOMENT APPROACHES TO APPROXIMATE INTERVAL ESTIMATION FOR A CERTAIN INTRACLASS CORRELATION COEFFICIENT

KELLY H. ZOU\textsuperscript{1,2}\* AND MICHAEL P. MCDERMOTT\textsuperscript{3}

\textsuperscript{1}Department of Health Care Policy, Harvard Medical School, Boston, Massachusetts 02115, U.S.A.
\textsuperscript{2}Department of Radiology, Brigham and Women's Hospital, Boston, Massachusetts 02115, U.S.A.
\textsuperscript{3}Department of Biostatistics, University of Rochester, Rochester, New York 14642, U.S.A.

SUMMARY

We consider the problem of constructing a confidence interval for the intraclass correlation coefficient in an interrater reliability study when the raters are assumed to be randomly selected from a population of raters. A Monte Carlo simulation study is conducted to investigate the true coverage probabilities of the commonly used intervals proposed by Fleiss and Shrout, which rely on Satterthwaite's two-moment approximation. These intervals are shown to be substantially anticonservative in certain cases. We propose intervals based on higher-moment approximations obtained using the Pearson system of distributions. The modified intervals are more conservative and generally more satisfactory than those obtained by the two-moment approximation. The competing methods are illustrated using data from a study of the natural history of facioscapulohumeral muscular dystrophy. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

The assessment of reliability of measurement is of great importance in medical research, where measurement error may have serious untoward consequences.\textsuperscript{1} There are various designs for reliability studies. For example, in test-retest or intrarater reliability studies, a single rater performs repeated assessments on a group of subjects. Alternatively, in interrater reliability studies, many raters each perform, typically, a single assessment on a group of subjects. This paper focuses on the problem of estimating the interrater reliability of an interval measurement.

Suppose that each of a random sample of $I$ raters independently rates each of a random sample of $J$ subjects, and that $Y_{ij}$, the rating by the $i$th rater on the $j$th subject, may be represented as

$$Y_{ij} = \mu + r_i + t_j + e_{ij} \text{ (} i = 1, \ldots, I; j = 1, \ldots, J\text{)}$$

where $\mu$ is the grand mean of the error-free measurements $Y$ in the population of interest; $r_i$ reflects the effect of the $i$th rater, which is assumed to be normally distributed with mean 0 and variance $\sigma_r^2$; $t_j$ is the effect of the $j$th subject, which is assumed to be normally distributed with
mean 0 and variance $\sigma_t^2$; and $e_{ij}$ represents the random error associated with this rating, which is also assumed to be normally distributed with mean 0 and variance $\sigma_e^2$. All random variables $\{t_i, t_j, e_{ij}\}$ are assumed to be mutually independent. Interrater reliability is typically quantified using the intraclass correlation coefficient, which may be interpreted as the proportion of the variability in the observed scores that can be accounted for by the subject-to-subject variability in the true (unobserved) scores:

$$\rho = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_r^2 + \sigma_e^2}.$$  \hspace{1cm} (2)

The analysis of variance table corresponding to the two-way random effects model (1) is given in Table I. From this table, the unbiased estimates of $\sigma_t^2$, $\sigma_r^2$ and $\sigma_e^2$ are easily seen to be $s_t^2 = (\text{SMS} - \text{EMS})/I$, $s_r^2 = (\text{RMS} - \text{EMS})/J$ and $s_e^2 = \text{EMS}$, respectively. Hence, Rajaratnam and Bartko proposed the following estimate of $\rho$:

$$\hat{\rho} = \frac{s_t^2}{s_t^2 + s_r^2 + s_e^2} = \frac{J(\text{SMS} - \text{EMS})}{J \times \text{SMS} + I \times \text{RMS} + (I - I - J)\text{EMS}}.$$  \hspace{1cm} (3)

Note that $\hat{\rho}$ may take negative values and in general is a biased estimate of $\rho$. However, the bias decreases as both $I$ and $J$ increase.

Fleiss and Shrout proposed a method for obtaining approximate confidence intervals for $\rho$ based on Satterthwaite’s two-moment approximation; this method is reviewed in Section 2. Alternative methods based on higher-moment approximations are introduced in Section 3. In Section 4, we present the results of a simulation study which reveals that the true coverage probabilities of the intervals based on the two-moment approximation tend to be conservative. We also propose some 'hybrid' methods which are generally more satisfactory in terms of producing intervals that have coverage probabilities closer to the nominal confidence coefficient. In Section 5, we illustrate the competing methods using data from a study of the natural history of facioscapulohumeral muscular dystrophy. Discussion and concluding remarks are provided in Section 6.
2. THE EXISTING TWO-MOMENT APPROACH

Consider testing $H_0: \rho = \rho_0$ versus $H_1: \rho > \rho_0$, where $\rho$ is given in (2). Since the distribution of $\hat{\rho}$ is difficult to obtain in general, a monotone function of $\hat{\rho}$ may be a more desirable test statistic. If the distribution of the test statistic is known under the null hypothesis, the appropriate testing procedure may easily be conducted for any given significance level. The ratio of SMS and $W$, a specified linear combination of RMS and EMS, serves such a purpose,\(^4\) where

$$W = \frac{I \rho}{J(1 - \rho)} \text{RMS} + \left\{1 + \frac{I(J - 1)\rho}{J(1 - \rho)}\right\} \text{EMS}.$$ 

Unfortunately the exact distribution of $W$, and hence of $F' = \text{SMS}/W$, is complicated. However, the problem of approximating the distribution of a linear combination of two mean squares has been addressed extensively in the literature, for example, in the Behrens–Fisher problem.\(^6\)

In particular, let $W_1$, $W_2$ and $W_3$ be three mutually independent random variables such that $v_k W_k / \theta_k \sim \chi_k^2 \ (k = 1, 2, 3)$. Suppose that it is of interest to approximate the distribution of $W = a_1 W_1 + a_2 W_2$, where the constants $a_1$ and $a_2$ are chosen so that the expectation of $W$ is equal to the expectation of the random variable $W_3$. Satterthwaite\(^5\) provided a solution to this problem by matching the first two moments of the exact distribution of $v_k W_k / \theta$ with those of a chi-square distribution with $v_k$ degrees of freedom, where $\theta = E[W] = a_1 \theta_1 + a_2 \theta_2$. With simple algebra, the approximate degrees of freedom $v_k$ is found to be

$$v_k = \frac{(a_1 u + a_2)^2}{a_1^2 u^2 + a_2^2} \frac{v_1}{v_2}$$

where $u = \theta_1 / \theta_2$.

Fleiss and Shrout\(^4\) applied Satterthwaite’s approximation to the above testing problem regarding the intraclass correlation coefficient. Specifically

$$W_1 = \text{RMS}, \quad W_2 = \text{EMS}, \quad W_3 = \text{SMS}$$

$$\theta_1 = \sigma_{\epsilon}^2 + J \sigma_{\sigma}^2, \quad \theta_2 = \sigma_{\epsilon}^2, \quad \theta_3 = \sigma_{\epsilon}^2 + I \sigma_{\sigma}^2$$

$$v_1 = I - 1, \quad v_2 = (I - 1)(J - 1), \quad v_3 = J - 1$$

$$a_1 = \frac{I \rho}{J(1 - \rho)}, \quad a_2 = 1 + \frac{I(J - 1)\rho}{J(1 - \rho)}.$$ \(4\)

Note that SMS and $W$ are independent and have the same expectation, therefore $F' = \text{SMS}/W$ is approximately distributed as Snedecor’s $F$ with $J - 1$ and $v_k$ degrees of freedom. In practice, the unknown quantities $\rho$, $\sigma_{\epsilon}^2$ and $\sigma_{\sigma}^2$ are replaced by their estimates $\hat{\rho}$, $\hat{s}_{\epsilon}^2$ and $\hat{s}_{\sigma}^2$, respectively, to obtain the estimated approximate degrees of freedom:

$$v_k = \frac{(I - 1)(J - 1)[I \hat{\rho} \hat{u} + J \{1 + (I - 1)\hat{\rho}\} - I \hat{\rho}]^2}{(J - 1)I^2 \hat{s}_{\epsilon}^2 \hat{u}^2 + [J \{1 + (I - 1)\hat{\rho}\} - I \hat{\rho}]^2}$$

where $\hat{u} = \text{RMS}/\text{EMS}$.

An approximate test of $H_0: \rho = \rho_0$ may then be conducted based on the $F$-distribution with $J - 1$ and $v_k$ degrees of freedom. Alternatively, the hypothesis test may be inverted to obtain
a 100(1 − α) per cent lower confidence bound for ρ:

\[ \hat{\rho}_L = \frac{J(SMS - F^* \times EMS)}{F^* \{I \times RMS + (IJ - I - J) \times EMS\} + J \times SMS} \]  

where \( F^* \) is the upper 100(1 − α) percentile of the \( F \) distribution with \( J - 1 \) and \( v_8 \) degrees of freedom.

3. HIGHER-MOMENT APPROACHES

We now develop some alternative methods under the rationale that a better approximation may arise by using higher-order moments when constructing an approximate confidence interval for \( \rho \). We use the first four standardized moments of the exact distribution of a linear combination of two mean squares to obtain a chi-square approximation. The following lemma is needed for this development:

**Lemma 1.** Let \( W_1 \) and \( W_2 \) be independent random variables such that \( \frac{v_k W_k}{\theta_k} \sim \chi^2_{\nu_k} (k = 1, 2) \). Let \( W = a_1 W_1 + a_2 W_2 \), where \( a_1 \) and \( a_2 \) are arbitrary real constants. Then

\[
E(W) = v_1 \gamma_1 + v_2 \gamma_2
\]

\[
E(W^2) = (v_1 + 2) \gamma_1^2 + (v_2 + 2) \gamma_2^2 + 2v_1 v_2 \gamma_1 \gamma_2
\]

\[
E(W^3) = v_1 (v_1 + 2)(v_1 + 4) \gamma_1^3 + 3v_1 v_2 (v_1 + 2) \gamma_1^2 \gamma_2 + 3v_1 v_2 (v_2 + 2) \gamma_1 \gamma_2^2 + v_2 (v_2 + 2)(v_2 + 4) \gamma_2^3
\]

\[
E(W^4) = v_1 (v_1 + 2)(v_1 + 4)(v_1 + 6) \gamma_1^4 + 4v_1 v_2 (v_1 + 2)(v_1 + 4) \gamma_1^3 \gamma_2 + 6v_1 v_2 (v_1 + 2)(v_2 + 2) \gamma_1^2 \gamma_2^2 + 4v_1 v_2 (v_2 + 2)(v_2 + 4) \gamma_1 \gamma_2^3 + v_2 (v_2 + 2)(v_2 + 4)(v_2 + 6) \gamma_2^4
\]

where \( \gamma_k = a_k \theta_k / v_k (k = 1, 2) \).

Lemma 1 may easily be proved by using the moment-generating functions of \( W_1 \) and \( W_2 \). The distribution of \( W \) may be approximated by a Pearson distribution having identical third and fourth standardized moments. It is known that probabilities involving general type III distributions in the Pearson system may always be expressed in terms of probabilities involving the chi-square distribution, therefore the ‘closest’ type III distribution to the above Pearson distribution may be chosen as the approximate distribution of \( W \).

By Lemma 1, the standardized third and fourth moments of the distribution of \( W \) are given by

\[ \beta_1^{1/2} = \frac{\mu_3}{\mu_2^{3/2}} = \left\{ 8 \left( \frac{a_1^3 u^3}{v_1^3} + \frac{a_2^3 u^3}{v_2^3} \right) \right\}^{1/2} \left( \frac{a_1^2 u^2}{v_1} + \frac{a_2^2 u^2}{v_2} \right)^{3/2} \]  

\[ \beta_2 = \frac{\mu_4}{\mu_2^2} = \left\{ 3a_1^4 u^4 (v_1 + 4) + 6a_1^2 a_2^2 u^2 v_1 v_2 + 3a_2^4 (v_2 + 4) \right\} \left( \frac{a_1^2 u^2}{v_1} + \frac{a_2^2 u^2}{v_2} \right)^2 \]  

respectively. The ‘closest’ type III distribution to the Pearson distribution having these standardized moments may be obtained by projecting \((\beta_1, \beta_2)\) onto the type III line \( 2\beta_2 - 3\beta_1 - 6 = 0 \) (see Figure 1), resulting in \((\beta_1^*, \beta_2^*)\). The vertical projection results in a three-moment approximation to the distribution of \( W \), whereas the orthogonal projection results in a four-moment approximation.
The degrees of freedom for the chi-square distribution associated with \((\beta^*_1, \beta^*_2)\) is given by

\[
v = \frac{8}{\beta^*_1^2}.
\]

If the ‘closest’ type III distribution is defined as the vertical projection \((\beta^*_1, \beta^*_2)\), then

\[
v_T = \frac{8}{\beta^*_{1T}} = \frac{8}{\beta^*_1}.
\]

(8)

Alternatively, if the orthogonal projection \((\beta^*_1, \beta^*_2)\) is used, then

\[
v_F = \frac{8}{\beta^*_{1F}} = 8/(4\beta_1 + 6\beta_2 - 18)/13.
\]

(9)

In the context of constructing an approximate confidence interval for the intraclass correlation coefficient, \(a_1, a_2\) and \(u\) in (6) and (7) are replaced with their corresponding estimates

\[
\hat{a}_1 = \frac{I\hat{\rho}}{J(1-\hat{\rho})}, \quad \hat{a}_2 = 1 + \frac{I(J-1)\hat{\rho}}{J(1-\hat{\rho})}, \quad \hat{u} = \text{RMS/EMS}
\]

to obtain \(\hat{\beta}_1\) and \(\hat{\beta}_2\), where \(\hat{\rho}\) is given in (3). By (8) and (9) we obtain \(\hat{v}_T\) and \(\hat{v}_F\), the three- and four-moment estimates of the approximate degrees of freedom. For example

\[
\hat{v}_T = 8/\hat{\beta}_1 = \left(\frac{\hat{a}_1^2}{v_1} + \frac{\hat{a}_2^2}{v_2}\right)^3 \left(\frac{\hat{a}_1^3}{v_1^2} + \frac{\hat{a}_2^3}{v_2^2}\right)^2.
\]
Similarly, \( \hat{\nu}_T W/(a_1 \theta_1 + a_2 \theta_2) \) is approximately distributed as chi-square with \( \hat{\nu}_T \) degrees of freedom.

\[
\hat{\nu}_F = \frac{8}{\{4\hat{\beta}_1 + 6\hat{\beta}_2 - 18\}/13}
\]

\[
= 52\left(\frac{a_1^2 \hat{u}^2}{v_1} + \frac{a_2^2 \hat{u}^2}{v_2}\right)^3 \left\{\frac{\hat{u}^a \hat{u}^b}{v_1^a v_2^b} (9v_1 + 52) + \frac{9\hat{a}_1^4 \hat{a}_2^4 \hat{u}^4}{v_1^4 v_2^4} (3v_1 + 4) + \frac{32\hat{a}_1^3 \hat{a}_2^3 \hat{u}^3}{v_1^2 v_2^2} \right\}
\]

\[
+ \frac{9\hat{a}_1^2 \hat{a}_2^2 \hat{u}^2}{v_1 v_2^2} (3v_2 + 4) + \frac{\hat{a}_2^2}{v_2^2} (9v_2 + 52) - 9\}
\]

and \( \hat{\nu}_F W/(a_1 \theta_1 + a_2 \theta_2) \) is approximately distributed as chi-square with \( \hat{\nu}_F \) degrees of freedom.

An empirical study using Monte Carlo simulation was undertaken to examine the true coverage probabilities of the competing approximate lower confidence bounds for \( \rho \). The following values were fixed for this study: \( \rho = 0.60, 0.75 \) and 0.90; \( I = 3 \) and 5; \( J = 10, 25 \) and 50; \( z = 0.05 \) and 0.10; and \( \sigma_r^2 / \sigma_\rho^2 = 0.5, 1.0 \) and 4.0. Each replication of the experiment consisted of randomly generating three independent \( \chi^2 \) random variables with respective degrees of freedom \( I - 1, J - 1 \) and \( (I - 1)(J - 1) \). The pseudorandom numbers were generated using the IMSL FORTRAN subroutine library. We then calculated \( \hat{\rho}_L \), the lower confidence bound for \( \rho \), and determined whether or not \( \hat{\rho}_L < \rho \), that is, the interval ‘covered’ the true value of \( \rho \). The true coverage probabilities were estimated by the proportion of times \( \hat{\rho}_L \) was less than \( \rho \) in 50,000 replications of the experiment.

Selected results of this study are presented in Tables II and III. The two-moment intervals based on Satterthwaite’s approximation (S) tend to be anticonservative, whereas the intervals based on the three- (T) and four-moment (F) approaches tend to be conservative. For example, in the cases studied, the 90 per cent lower confidence bounds based on Fleiss and Shrout’s approach have true coverage probabilities as low as 75 per cent. On the other hand, the 90 per cent lower confidence bounds based on the three- and four-moment approaches have true coverage probabilities as high as 97 per cent. In particular, for fixed \( I \), the performance of the approximation tends to decline with increasing \( J \), and also with increasing \( \sigma_r^2 / \sigma_\rho^2 \). Also, for fixed \( J \), the performance of the two-moment approximation tends to improve with increasing \( I \), whereas the performance of the higher-moment approximations tends to decline with increasing \( I \).

Following Scariano and Davenport’s study of the Behrens–Fisher problem, we considered choosing a weighted average of \( \hat{\nu}_S \) and \( \hat{\nu}_T \) (or \( \hat{\nu}_F \)) as the denominator degrees of freedom for the approximate \( F \)-distribution. In particular, the hybrid approaches based on \( \hat{\nu}_T^* = p\hat{\nu}_S + (1 - p)\hat{\nu}_T \) and \( \hat{\nu}_F^* = p\hat{\nu}_S + (1 - p)\hat{\nu}_F \) were studied, where \( p \in (0, 1) \). Results for \( p = 0.1 \) are presented in Tables II and III. These hybrid methods (T* and F*) reduce the conservatism of the three- and four-moment approaches, but we found that placing more weight than \( p = 0.1 \) on \( \hat{\nu}_S \) made the hybrid approaches unacceptably anticonservative in some cases.

Copyright © 1999 John Wiley & Sons, Ltd.

Table II. Empirical coverage probabilities (multiplied by 1000) of approximate 90 per cent lower confidence bounds for $\rho$, based on 50,000 simulations

<table>
<thead>
<tr>
<th>$\sigma^2/\sigma_e^2$</th>
<th>0.5</th>
<th>1.0</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>I</td>
<td>J</td>
<td>S</td>
</tr>
<tr>
<td>0.60</td>
<td>3</td>
<td>10</td>
<td>896</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>868</td>
<td>914</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>840</td>
<td>912</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>897</td>
<td>928</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>882</td>
<td>938</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>862</td>
<td>950</td>
</tr>
<tr>
<td>0.75</td>
<td>3</td>
<td>10</td>
<td>888</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>863</td>
<td>910</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>830</td>
<td>905</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>894</td>
<td>925</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>880</td>
<td>940</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>856</td>
<td>949</td>
</tr>
<tr>
<td>0.90</td>
<td>3</td>
<td>10</td>
<td>888</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>857</td>
<td>907</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>820</td>
<td>900</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>894</td>
<td>927</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>874</td>
<td>940</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>854</td>
<td>948</td>
</tr>
</tbody>
</table>

S = two-moment approximation; T = three-moment approximation; F = four-moment approximation; T* = hybrid of S and T; F* = hybrid of S and F
Table III. Empirical coverage probabilities (multiplied by 1000) of approximate 95 per cent lower confidence bounds for $\rho$, based on 50,000 simulations

<table>
<thead>
<tr>
<th>$\sigma_i^2/\sigma_p^2$</th>
<th>0.5</th>
<th>1.0</th>
<th>4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$ I J</td>
<td>S</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>0.60 3 10</td>
<td>945</td>
<td>958</td>
<td>960</td>
</tr>
<tr>
<td>25 946</td>
<td>950</td>
<td>954</td>
<td>950</td>
</tr>
<tr>
<td>50 989</td>
<td>941</td>
<td>947</td>
<td>938</td>
</tr>
<tr>
<td>5 10 946</td>
<td>963</td>
<td>966</td>
<td>958</td>
</tr>
<tr>
<td>25 933</td>
<td>968</td>
<td>972</td>
<td>964</td>
</tr>
<tr>
<td>50 919</td>
<td>972</td>
<td>977</td>
<td>966</td>
</tr>
<tr>
<td>0.75 3 10</td>
<td>941</td>
<td>954</td>
<td>956</td>
</tr>
<tr>
<td>25 920</td>
<td>948</td>
<td>952</td>
<td>942</td>
</tr>
<tr>
<td>50 892</td>
<td>940</td>
<td>944</td>
<td>935</td>
</tr>
<tr>
<td>5 10 946</td>
<td>964</td>
<td>967</td>
<td>953</td>
</tr>
<tr>
<td>25 933</td>
<td>968</td>
<td>972</td>
<td>965</td>
</tr>
<tr>
<td>50 913</td>
<td>970</td>
<td>976</td>
<td>965</td>
</tr>
<tr>
<td>0.90 3 10</td>
<td>941</td>
<td>954</td>
<td>957</td>
</tr>
<tr>
<td>25 917</td>
<td>945</td>
<td>950</td>
<td>941</td>
</tr>
<tr>
<td>50 885</td>
<td>936</td>
<td>943</td>
<td>928</td>
</tr>
<tr>
<td>5 10 945</td>
<td>963</td>
<td>966</td>
<td>959</td>
</tr>
<tr>
<td>25 920</td>
<td>966</td>
<td>971</td>
<td>964</td>
</tr>
<tr>
<td>50 910</td>
<td>968</td>
<td>974</td>
<td>964</td>
</tr>
</tbody>
</table>

S = two-moment approximation; T = three-moment approximation; F = four-moment approximation; T* = hybrid of S and T; F* = hybrid of S and F
5. AN EXAMPLE

As part of a study of the natural history of facioscapulohumeral muscular dystrophy (FSHD), measurements of maximum voluntary isometric contraction (MVIC) were obtained in 18 muscles from each patient using quantitative myometry. The testing device consisted of an adjustable cuff attached by an inelastic strap to a linear force transducer. The testing involved putting the patient’s limb in a standardized position, attaching the cuff to the limb, and instructing the patient to pull on the strap with as much force as possible using only the targeted muscle group. There are many sources of possible measurement error in this type of strength testing, such as the evaluator’s initial positioning of the patient, the positioning of the adjustable cuff, possible substitution of other muscles in trying to generate force, and variability of effort on the part of the patient.

Since the study was performed in two different centres, a pilot interrater reliability study was performed for the two clinical evaluators. Seven patients were tested on two consecutive days, with the order of testing by evaluator 1 and evaluator 2 randomly determined for each patient. Table IV displays the resulting two-way analysis of variance table for the force in Newtons generated by the right quadricep (knee extension). It is easy to verify that $\hat{\rho} = 0.898$.

The approximate degrees of freedom and lower confidence bound for $\rho$ are provided in Table V for each of the five competing methods, with $p = 0.1$ chosen for the hybrid methods ($T^*$ and $F^*$).

6. DISCUSSION

The method based on Satterthwaite’s two-moment approximation is widely used for computing approximate confidence intervals for the intraclass correlation coefficient in an interrater reliability study. However, in a study that has relatively few raters and many subjects, a situation that may be most common in practice, this method produces intervals which are anticonservative and substantially so in some cases. The method may also be unsatisfactory in studies in which the rater component of variability in the ratings is expected to be large relative to the random error.
component. This may be the case, for example, if the measuring instrument under study had extremely good intrarater or test–retest reliability.

Our simulation results illustrate that for a fixed number of raters \( I \), the performance of the two-moment approximation becomes worse as the number of subjects \( J \) increases. We tried to examine this counterintuitive result analytically using moment generating functions since these are easily derived for the random variable \( W \) as well as for a chi-square random variable. However, computation of the limiting moment generating functions turned out to be intractable.

Instead, we performed several Monte Carlo experiments. First, we fixed \( \sigma_r^2 = \sigma_e^2 = 1 \) and \( \rho = 0.75 \). Next, we generated 50,000 random variables from the approximate and exact distributions of \( W \) and examined quantile-quantile plots for various values of \( I \) and \( J \) (data not shown). The results confirmed our observation that, for fixed \( I \), the agreement among the two distributions becomes worse as \( J \) increases. They also confirmed that, for fixed \( J \), the agreement improves as \( I \) increases. In general, for large values of \( I \) the agreement between the approximate and exact distributions is quite good. It is known that Satterthwaite’s approximate distribution of a linear combination of mean squares tends to normality as the degrees of freedom jointly approach infinity.\(^3\) The same is true of the exact distribution. The degrees of freedom involved in the mean squares used to compute \( W \) are \( (I - 1) \) and \( (I - 1)(J - 1) \), both of which depend on \( I \). This may explain our findings. The situation considered here is somewhat more complex, however, since the coefficients involved in the linear combination also depend on both \( I \) and \( J \).

The proposed methods based on higher-moment approximations were seen to yield more conservative intervals than those of two-moment method, however these intervals can be substantially conservative. It should be noted that while these methods utilize information from either three or four moments of the exact distribution of the appropriate linear combination of mean squares, these moments cannot in general be forced to be the same for the exact and approximate distributions. This is in contrast to the two-moment method based on Satterthwaite’s approximation, which actually equates the first two moments of the exact and approximate distributions.\(^4,5\) The rationale for using the higher-moment approaches is that using a larger amount of ‘approximate’ information may be better than using a smaller amount of ‘exact’ information. Therefore none of these methods seems to be uniformly better than the others, although the conservative nature of the higher-moment approaches may make them more desirable in practice. A ‘hybrid’ approach with relatively small \( p \) can be used to reduce this conservatism.

ACKNOWLEDGEMENTS

We acknowledge with thanks helpful comments from Professor Heng Li of the University of Rochester and careful review by the two anonymous referees.

REFERENCES

2. Rajaratnam, N. ‘Reliability formulas for independent decision data when reliability data are matched’, *Psychometrika*, 25, 261–271 (1960)