Multistage Evaluation of Measurement Error in a Reliability Study

Aiyi Liu*, Enrique F. Schisterman and Chengqing Wu

Division of Epidemiology, Statistics and Prevention Research, National Institute of Child Health and Human Development, NIH/DHHS, 6100 Executive Blvd., Rockville, MD 20852, U.S.A.

**email:* liua@mail.nih.gov

SUMMARY. We introduce sequential testing procedures for the planning and analysis of reliability studies to assess an exposure's measurement error. The designs allow repeated evaluation of reliability of the measurements and stop testing if early evidence shows the measurement error is within the level of tolerance. Methods are developed and critical values tabulated for a number of two-stage designs. The methods are exemplified using an example evaluating the reliability of biomarkers associated with oxidative stress.

KEY WORDS: Interim analysis; Intraclass correlation; Measurement error; Power and sample size.

1. Introduction

In epidemiologic studies evaluating the association between outcomes and exposures, measurements of exposures are often subject to error. Accurately assessing the amount of and adequately correcting such error is important since otherwise substantial bias can be introduced to the estimates of regression coefficients or relative risks, as well as

sample size and power of the study (Liu et al. 1978; Armstrong, White and Saracci, 1992; Carroll, Ruppert and Stefanski, 1995; Freedman et al., 2004).

Reliability studies are a common approach to assess the reproducibility of the measures of an exposure, i.e. how consistently the measurements of the exposure can be repeated on the same subject. By the nature of such studies, the exposure's values of the same subject are measured several times using the same instrument or different ones, either simultaneously or at several time points. Consistency of the measurements is often assessed using the so called intraclass correlation coefficient with larger values indicating higher level of consistency (Armstrong et al., 1992).

In many studies, measuring an exposure values can be very costly and the cost is driven high if repeated measurements are taken. The need to reduce study cost is clearly a motive to adopt the idea of sequential testing that is widely used in clinical trials (e.g. Pocock, 1977; O'Brien and Fleming, 1979; Lan and DeMets, 1983; Slud, 1984; Kim and DeMets, 1987; Tang, Geller and Pocock, 1993; Tan, Xiong, and Kutner, 1996; Whitehead, 1997; Jennison and Turnbull, 2000) to the design and analysis of reliability studies. Furthermore, when there are several exposures of interest, it is desirable to stop evaluation of an exposure and reallocate the resources to others if early evidence shows the exposure to have small (or large) measurement error.

In this paper, we propose to sequentially evaluate the measurements' reliability of an exposure. If at any early stage the data shows substantial consistency in the measurements, then the evaluation process stops. Otherwise, it continues to the next stage till a conclusion regarding the measurement error can be made. Such designs are cost-effective as compared to the traditional one-stage designs (e.g. Kraemer and Korner, 1976; Donner and Eliasziw, 1987). In Section 2, we describe a motivating example, review briefly the sample size and power calculation for one-stage reliability studies,

and then introduce a general sequential procedure for testing hypothesis concerning the measurement error. In Section 3 we propose methods to derive critical values, sample size and power, and present simulation results for a number of two-stage designs using Lan and DeMets error spending appraoch, and revisit the motivating example. Summary and discussions are given in Section 4.

2. A Motivating Example and the Multistage Procedure

2.1 *Example: A study of oxidative biomarkers*

Studies have suggested that oxidative stress might be implicated in the risk of human infertility. However, mechanisms by which oxidative stress may be associated with female infertility are not completely understood. Although it is known that micronutrient antioxidants, hormones and enzymatic antioxidants neutralize oxygen-free radicals and inhibit oxidation, lack of knowledge on the association and interrelation of these antioxidants with oxidative stress levels across the menstrual cycle in human females has hindered the use of these therapies to reduce the risk of infertility. For this reason, the BioCycle study, a longitudinal study to assess the effects of endogenous hormones (i.e. estrogen and progesterone) on biomarkers of oxidative stress and antioxidant status during the menstrual cycle, was initiated at the National Institute of Child Health and Human Development (NICHD) of the National Institutes of Health (NIH). One part of the first phase of the study is to enroll a number of 10 women to assess the variation in measures of F2 Isoprostanes, an important oxidative stress biomarker, during various phases of the menstrual cycle. It was planned that at each specific time point within a menstrual cycle, the F2 Isoprostanes values will be measured simultaneously three times, which is used to assess the consistency of the measurements.

This is a typical study of reliability. Let ρ denote the intraclass correlation co-

efficient (to be defined below) of F2 Isoprostanes measures at a selected time point. Then the null hypothesis to be tested in this study is $H_0 : \rho \leq \rho_0 (= 0.5)$ with level of significance to be $\alpha (= 0.05)$. Since each assay costs about \$130, cost reduction was considered when planning for the study.

2.2 *One-stage test*

In general, an analysis of variance (ANOVA) approach can be employed to analyze a reliability study such as the F2 Isoprostanes example described above (repeated measurements at a time point). Let X_{ij} be the jth ($j = 1, \ldots, p$) measurement from the ith ($i = 1, \ldots n$) subject, where the n subjects form the independent experimental units of the study. Then the one-way ANOVA model is

$$X_{ij} = \mu + u_i + \epsilon_{ij}, \tag{1}$$

where the fixed effect μ is the grand mean of all measurements, the random effects u_i, reflecting the variation among the subjects, are normally distributed with mean 0 and variance σ_u^2, and the measurement error ϵ_{ij} are normally distributed with mean 0 and variance σ_ϵ^2 whose magnitude reflects the variation among the measurements within a subject. Further we assume that u_i are independent of ϵ_{ij}. Note that these assumptions lead to $Cov(X_{ij}, X_{i'j'}) = 0$ if $i \neq i'$, $Cov(X_{ij}, X_{ij'}) = \sigma_u^2$ if $j \neq j'$, and $Var(X_{ij}) = \sigma_u^2 + \sigma_\epsilon^2$.

A popular measure of consistency of the within-subject measurements is the intra-class correlation coefficient defined as

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2}\,. \tag{2}$$

This is the correlation coefficient between two measurements from the same subject. Larger values of ρ indicates higher coherence among measurements from the same

subject since then the within-subject error is relatively smaller as compared to the between subject error; perfect consistency occurs when $\rho = 1$ (then $\sigma_\epsilon = 0$).

Define $S_W^2 = \sum_{i=1}^n \sum_{j=1}^p (X_{ij} - \bar{X}_{i.})^2$, the within sum of squares, and $S_B^2 = \sum_{i=1}^n p(\bar{X}_{i.} - \bar{X}_{..})^2$, the between sum of squares, where $\bar{X}_{i.} = \sum_{j=1}^p X_{ij}/p$, the within-subject average, and $\bar{X}_{..} = \sum_{i=1}^n \sum_{j=1}^p X_{ij}/(np)$, the overall average. Then (e.g. Rao 1972) S_W^2 and S_B^2 are independent and

$$S_W^2 \sim \sigma_\epsilon^2 \chi^2_{n(p-1)},\ S_B^2 \sim (\sigma_\epsilon^2 + p\sigma_u^2)\chi^2_{n-1}, \tag{3}$$

where χ_m^2 stands for the (central) chi-square distribution with m degrees of freedom. Thus $E(S_W^2) = n(p-1)\sigma_\epsilon^2$, and $E(S_B^2) = (n-1)(\sigma_\epsilon^2 + p\sigma_u^2)$, yielding an estimate of ρ (the sample intraclass correlation coefficient):

$$\hat{\rho} = \frac{n(p-1)F - (n-1)}{n(p-1)F + (n-1)(p-1)}, \tag{4}$$

where $F = S_B^2/S_W^2$. Note that $\hat{\rho}$ is strictly increasing in F. Further, from (3), we have

$$F \sim \frac{(n-1)(1+(p-1)\rho)}{n(p-1)(1-\rho)} F_{n-1,n(p-1)}, \tag{5}$$

with F_{m_1,m_2} being a (central) F-distribution with degrees of freedom m_1 and m_2.

With these developments, the one-stage testing procedure with level of significance α rejects the null hypothesis $H_0 :\ \rho \leq \rho_0$ if $\hat{\rho} > c$ for such $0 < c < 1$ that $P_{\rho=\rho_0}(\hat{\rho} > c) = \alpha$. To compute c, note that $\hat{\rho} > c$ if and only if $F > (n-1)(1+(p-1)c)/(n(p-1)(1-c))$. Setting the probability of the latter event to α and utilizing (5), we find, through straightforward algebraic manipulation, that

$$c = \frac{(1+(p-1)\rho_0)F^{-1}_{n-1,n(p-1)}(1-\alpha) - (1-\rho_0)}{(p-1)(1-\rho_0) + (1+(p-1)\rho_0)F^{-1}_{n-1,n(p-1)}(1-\alpha)}. \tag{6}$$

The power of such test at $\rho > \rho_0$ is thus given by

$$P_\rho(\hat{\rho} > c) = 1 - F_{n-1,n(p-1)}\left(\frac{(1-\rho)(1+(p-1)\rho_0)}{(1-\rho_0)(1+(p-1)\rho)} F^{-1}_{n-1,n(p-1)}(1-\alpha)\right). \tag{7}$$

Setting (7) to the desired power at an alternative value of ρ and numerically solving the equation then yields the sample size needed for the study.

2.2 *Multistage test*

Suppose the subjects are taken into groups. Denote by g_k the number of subjects in the kth group, and $n_k = g_1 + g_2 + \ldots + g_k$ the cumulative sample size (number of subjects) of the first k groups. Write $\hat{\rho}_k$ as the estimate of the intraclass correlation coefficient ρ computed from data observed up to the kth group. Let K be the pre-specified maximum number of tests planned for the study. Then a K-stage testing procedure for $H_0 : \rho < \rho_0$ is carried out as follows. First, p repeated measurements are taken from each of the g_1 subjects in the first group, and $\hat{\rho}_1$ is computed. If $\hat{\rho}_1 > c_1$, then the test stops and H_0 is rejected. Otherwise, measurements from the g_2 subjects in the second group are taken, and $\hat{\rho}_2$ is computed based on data from all n_2 subjects. In general, at each stage $k = 1, \ldots, K-1$, with critical values c_k, the sampling process stops with rejection of H_0 if

$$\hat{\rho}_i \leq c_i, i \leq k-1, \text{ and } \quad \hat{\rho}_k > c_k \,. \tag{8}$$

Otherwise, the test continues to stage $k+1$; if it did not stop at any early stage $k < K$, then it will stop at the Kth stage and reject H_0 if $\hat{\rho}_K > c_K$.

Such "group" sequential testing concepts have become standard practice in clinical trials. The power function of the test is

$$\beta(\rho) = \sum_{k=1}^{K} P(\hat{\rho}_i \leq c_i, i \leq k-1, \text{ and } \quad \hat{\rho}_k > c_k) \,, \tag{9}$$

where the critical values c_k are chosen so that the type I error of the test, $\beta(0)$, is α and the power at $\rho_1 > \rho_0$, $\beta(\rho_1)$, is $1-\beta$, where α and β are pre-specified rates of type I and type II error, respectively. In a clinical trial setting where testing is focused on a normal mean or a Bernoulli probability, a considerable number of methods have been

proposed to compute these critical values; see Jennison and Turnbull (2000). In the next section we will focus on two-stage designs and explore Lan and DeMets (1983) error speeding approach which allocates the overall type I error α to each stage till the overall error "is spent". Once an error spending function is chosen, the kth summand in (9), i.e. the error spent at the kth stage, will be given by the error spending function.

3. Two-stage Designs and Simulation Technique

3.1 *Determination of critical values*

We write S^2_{1W}, S^2_{1B} and S^2_{2W}, S^2_{2B} to be the within- and between- sum of squares computed at the first and second stage, with corresponding estimates of intraclass correlation coefficient $\hat{\rho}_1$ and $\hat{\rho}_2$, respectively. Suppose the type I error to be spent at stage 1 is α_1. Then from (9) the error requirements (overall type I error α and power $1-\beta$ at ρ_1) mandate that

$$P_{\rho_0}(\hat{\rho}_1 > c_1) = \alpha_1\,, \quad P_{\rho_0}(\hat{\rho}_1 \le c_1, \hat{\rho}_2 > c_2) = \alpha - \alpha_1\,, \tag{10}$$

and

$$P_{\rho_1}(\hat{\rho}_1 > c_1) + P_{\rho_1}(\hat{\rho}_1 \le c_1, \hat{\rho}_2 > c_2) = 1 - \beta\,. \tag{11}$$

These three equations together determine the critical values and the sample size assuming the allocation ratio, n_1/n_2, of subjects to the first group is specified. Recalling (6), the first equation of (10) yields

$$c_1 = \frac{(1+(p-1)\rho_0)F^{-1}_{n_1-1,n_1(p-1)}(1-\alpha_1) - (1-\rho_0)}{(p-1)(1-\rho_0) + (1+(p-1)\rho_0)F^{-1}_{n_1-1,n_1(p-1)}(1-\alpha_1)}. \tag{12}$$

Thus the first summand of (11) is given by (7) with substitution of (ρ, c, n) for (ρ_1, c_1, n_1). The second equation of (10) and the second summand of (11) involve evaluation of $\beta_2(\rho) = P_\rho(\hat{\rho}_1 \le c_1, \hat{\rho}_2 > c_2)$ for some ρ. This requires knowledge of the joint distribution of $\hat{\rho}_1$ and $\hat{\rho}_2$ for which no closed form is available. The following

results (proof given in the appendix) provide a simple formula for the distribution which substantially simplifies computation and simulation concerning the error probabilities. Let U_1, U_2, U_3 and U_4 be independent random variables having χ^2 distributions with degrees of freedom $n_1 - 1$, $n_1(p-1)$, g_2, and $g_2(p-1)$, respectively and define

$$b(n, c; \rho) = \frac{(1-\rho)(n-1)(1+c(p-1))}{n(p-1)(1-c)(1+(p-1)\rho)}. \tag{13}$$

Then

$$\begin{aligned} P\left(\hat{\rho}_1 \leq c_1,\ \hat{\rho}_2 > c_2\right) &= P\Big(U_1 \leq b_1 U_2,\ U_1 + U_3 > b_2(U_2 + U_4)\Big) \\ &= \int_0^{b_1} \left(\int_\Delta d\chi_{n_1(p-1)}(u_1) d\chi_{g_2}(u_3) d\chi_{g_2(p-1)}(u_4) \right) dF_{n_1-1, n_1(p-1)} \left(\frac{n_1(p-1)}{n_1 - 1}\, y \right), \end{aligned} \tag{14}$$

where $b_1 = b(n_1, c_1; \rho)$, $b_2 = b(n_2, c_2; \rho)$, and $\Delta = \{(u_2, u_3, u_4) : (y - b_2)u_2 + u_3 > b_2 u_4\}$.

The above formulas, along with (10) and (11), provide two options to compute the critical values and the power (or sample size) of the test. One is to use multi-variable integrations in the second equation. This, however, requires computational capacity and accuracy for multiple integrals and often proves to be difficult. Alternatively, one can use the first equation to simulate data from the four independent χ^2 distributions and then find the critical values subject to error requirements in (10) and (11). This approach is computationally straightforward, simpler and faster than simulation from multivariate normal distributions. Providing the number of simulations is sufficiently large, it in general yields accurate results. We use this latter approach to compute critical values for various two-stage designs; see more below.

3.2 *Power analysis of designs with given sample size: The F2 Isoprostanes example revisited*

In many studies the sample size n_2 is given, and analysis of power at various alternatives is desirable. These designs are particularly employed when the number of subjects that can be studied is limited (e.g. due to budget constraints).

Computation and simulation related to these designs is relatively less extensive. With n_1, α and α_1 also given, the first critical value c_1 is given by (12). The second critical value c_2 can be determined via simulation based on the four independent χ^2 variables $U = (U_1, U_2, U_3, U_4)$ in (14). For a range of values of c_2, with a total of N runs of U, the empirical probability of the left side of the second equation of (10) can be obtained and the required c_2 value is the one that gives empirical probability equal or close to $\alpha - \alpha_1$. Once c_1 and c_2 are determined, the first summand of (11) can be computed using (7) and the second summand is obtained via simulation using the U-observations; addition of these two summands then yields the power of the test at an alternative value. To narrow the search range for c_2, we start with the one-stage critical value computed based on (6) with $n = n_2$, and then expand to the neighborhood till the desired value of c_2 is found.

We now turn to the F2 Isoprostanes example described in Section 2.1. The interim analysis was planned when half of the subjects supply their blood samples and the values of F2 Isoprostanes are measured. The interim error α_1 was set to be $\alpha/2$ based on an error spending function of Kim and DeMets (1987); see below. Thus $\rho_0 = 0.5$, $n_1 = n_2/2 = 5$, and $\alpha_1 = \alpha/2 = 0.025$. With these design parameters, we simulated $N = 100,000$ runs of random vector U, and found $c_1 = 0.8493$, and $c_2 = 0.7593$. The power of the test at $\rho_1 = 0.80, 0.85, 0.90$ is 0.65, 0.81 and 0.94, respectively. The average sample size at these alternative values is 9, 8, and 7, respectively, reflecting the cost-effective nature of the design.

3.3 *Computation of sample size and simulation results*

In other studies the sample size n_2 needs to be determined based on pre-specified significance level α and power $1-\beta$ at an alternative $\rho = \rho_1$. These designs often occur when it is feasible to study a relatively larger number of subjects. We demonstrate

below how to find the critical values c_1, c_2 and the required total sample size n_2 via Monto Carlo simulation. To this end, we set $n_1 = n_2/2$. For illustration we use error spending functions of Kim and DeMets (1987), $\alpha_1 = \alpha/2^r$. (The F2 Isoprostanes example above used $r = 1$.) Possible values of design parameters to be considered are $\alpha = \{0.025, 0.05\}$, for testing null value $\rho_0 = \{0.5, 0.6\}$, power of the test $1 - \beta = \{0.8, 0.85, 0.9\}$ at $\rho = \rho_1 = \{0.70, 0.75, 0.80, 0.85, 0.90, 0.95\}$, with $r = \{1, 1.5, 2\}$. Note that the null values chosen indicate moderate consistency in measurements, according to Landis and Koch (1977). We consider $p = 2, 3, 4$ up to 10 replicates, respectively.

For each set of design parameters $(\alpha, \beta, r, \rho_0, \rho_1)$, we search a potential range of (c_1, c_2, n_2) to find a combination that meets the error requirements. For each combination of (c_1, c_2, n_2), c_1 can be determined either via simulation or by (12). A total number of $N = 100,000$ random observations from the random vector $U = (U_1, U_2, U_3, U_4)$ of the four independent χ^2 variables are drawn, and the empirical probabilities in (10) and (11) are computed. The combination that satisfies error requirements in (10) and (11) is thus the desired solution to the design. Again, to help narrow the range, we first limit the values of c_2 and n_2 close to that required in a one-stage test by (6) and (7), and then gradually expand the range till a solution is found.

Table 1 presents simulation results of critical values and sample size required, for a selected number of designs with $r = 1, 2$. (The complete tabulation and the program codes are available upon request.) It is noticed that setting $r = 2$ yields more conservative stopping boundaries (less likely to stop early) than $r = 1$, and thus requires smaller sample size. For each design, the final critical value c_2 is smaller than the interim critical value c_1, partly reflecting the increase of sample size. Again, the average sample size (ASN) is smaller than the fixed sample size (required in the one-stage procedure).

Table 1

Sample size tabulation for a number of two-stage designs

Put Table 1 here

4. Discussion

In this article we introduced sequential methods for evaluation of measurement error, and presented stopping boundary values for a number of two-stage designs. To our best knowledge there has been no previous work in the literature advocating sequential testing procedures to evaluate measurement errors. In addition to high cost of measuring the exposure, in some situations ethical consideration may also provide another motive to use these designs since some procedures such as taking bone marrow samples may be very painful to the study subjects and thus sampling process should be stopped if early evidence shows the measurement error is within the level of tolerance.

Unlike the sequential testing procedures developed specifically for clinical trials based on testing a normal mean or a Bernoulli probability, the methods in this article requires computation of the joint distributions of a series of sample intraclass correlation coefficients, or equivalently, a series of F-statistics. It is worth noting that these distributions involve only the correlation coefficient and are independent of the means and variances of the repeated measurements. Moreover, under both null and alternative hypotheses, computation of error probabilities requires only central F and χ^2 distributions. In contrast sequential F-tests concerning the mean vector of a multivariate normal variable involve non-central F and χ^2 distributions under the alternative hypotheses; see Jennison and Turnbull (2000, chapter 15).

Future research is needed to develop and compare various sequential procedures and to propose methods for post-test analysis such as computation of p-values and point and interval estimation of the intraclass correlation coefficient.

References

Liu, K., Stammler, J., Dyer, A., McKeever, J., and McKeever, P. (1978). Statistical methods to assess ans minimize the role of intr-individual variability in obscring the relationship between dietary lipids and serum cholesterol. Journal of Chronic Diseases **31**, 399-418.

Armstrong, B. K., White, E., and Saracci, R. (1992). *Principles of Exposure Measurement in Epidemiology.* New York: Oxford University Press.

Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models.* Boca Raton: Chapman and Hall.

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

Freedman, L. S., Fainberg, V., Kipnis, V., Midthune, D. and Carroll, R. J. (2004). A new method for dealing with measurement error in explanatory variables of regression models. *Biometrics* **60**, 172-181.

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials.* New York: Chapman and Hall/CRC.

Kim, K. and DeMets, D. L. (1987). Design and analysis of group sequential tests based on type I error spending rate function. *Biometrika* **74**, 149-154.

Kraemer, H. C. and Korner, A. F. (1979). Statistical alternatives in assessing reliability, consistency and individual dierences for quantitative measures: applications to behavioural measures of neonates. *Psychological Bulletin* **83**, 914-921.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**, 159-174.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659-663.

O'Brien, P. C. and Fleming. T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549-556.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191-100.

Rao, C. R. (1972). *Linear Statistical Inference and Its Applications.* New York: Wiley.

Slud, E. V. (1984). Sequential linear rank tests for two-sample censored survival data. *Annals of Statistics* **12**, 551-571.

Tan, M., Xiong, X. and Kutner, M. H. (1998). Clinical trial designs based on sequential conditional probability ratio tests and reverse stochastic curtailing. *Biometrics* **54**, 684-697.

Tang, D., Geller, N. L. and Pocock, S. J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics* **49**, 23-30.

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*, Revised 2nd ed. New York: Wiley.

Appendix

Proof of (14)

Define

$$U_1 = \frac{1}{\sigma_\epsilon^2 + p\sigma_u^2} S_{1B}^2,\ U_2 = \frac{1}{\sigma_\epsilon^2} S_{1W}^2,$$
$$U_3 = \frac{1}{\sigma_\epsilon^2 + p\sigma_u^2} (S_{2B}^2 - S_{1B}^2),\ U_4 = \frac{1}{\sigma_\epsilon^2} (S_{2W}^2 - S_{1W}^2).$$

It follows from (3) that $U_1 \sim \chi^2_{n_1-1}$, $U_2 \sim \chi^2_{n_1(p-1)}$, and U_1 and U_2 are independent. For U_4, because

$$S_{2W}^2 = \sum_{i=1}^{n_2} \left\{ \sum_{j=1}^{p} (X_{ij} - \bar{X}_{i.})^2 \right\} = S_{1W}^2 + \sum_{i=n_1+1}^{n_2} \left\{ \sum_{j=1}^{p} (X_{ij} - \bar{X}_{i.})^2 \right\},$$

hence $U_4 \sim \chi^2_{g_2}$ and is independent of U_1 and U_2. To find the distribution of U_3, define $\bar{X}_{1..} = \sum_{i=1}^{n_1} \sum_{j=1}^{p} X_{ij}/(n_1 p)$, $\bar{X}_{2..} = \sum_{i=1}^{n_2} \sum_{j=1}^{p} X_{ij}/(n_2 p)$, $\bar{X}_{21..} = \sum_{i=n_1+1}^{n_2} \sum_{j=1}^{p} X_{ij}/(g_2 p)$, and $S_{21B}^2 = \sum_{i=n_1+1}^{n_2} (\bar{X}_{i.} - \bar{X}_{21..})^2$. The first two terms are respectively the grand sample means at the first and second stage, and the last two terms are respectively the grand mean and between-subject sum of squares of observations in the second group alone. Then through straightforward (but tedious) algebraic manipulations, we obtain

$$U_3 = \frac{1}{\sigma_\epsilon^2 + p\sigma_u^2} S_{21B}^2 + \frac{1}{\sigma_\epsilon^2 + p\sigma_u^2} \frac{n_1 n_2 p}{g_2} \left(\bar{X}_{2..} - \bar{X}_{1..} \right)^2.$$

Again from Rao (1972) on distribution of quadratic forms of multivariate normal variables, the two summands in the equation above are independent, follow χ^2 distributions with degrees of freedom $g_2 - 1$ and 1, respectively, and both are independent of S_{1B}^2. Thus U_3 also follows a χ^2 distribution with degrees of freedom g_2.

The first equation of (14) hence follows because $\hat{\rho}_1 \leq c_1$ and $\hat{\rho}_2 > c_2$ are the same as $U_1/U_2 \leq b(n_1, c_1; \rho)$ and $(U_1 + U_3)/(U_2 + U_4) > b(n_2, c_2; \rho)$. The second equation of (14) is derived by using conditioning arguments.

Below is Table 1

alpha	beta	r	rho_0	rho_1	n_1	c_1	c_2	ASN	fixed size
					p=2				
0.05	0.1	1	0.5	0.7	46	0.6874	0.6284	66.05	83
0.05	0.1	2	0.5	0.7	44	0.7133	0.6250	69.06	83
0.025	0.1	1	0.5	0.7	55	0.6943	0.6354	80.74	102
0.025	0.1	2	0.5	0.7	53	0.7156	0.6319	84.25	102
0.05	0.2	1	0.5	0.7	33	0.7155	0.6493	51.82	61
0.05	0.2	2	0.5	0.7	32	0.7425	0.6423	54.08	61
0.025	0.2	1	0.5	0.7	43	0.7153	0.6528	67.91	78
0.025	0.2	2	0.5	0.7	40	0.7414	0.6510	68.23	78
0.05	0.1	1	0.5	0.8	16	0.7864	0.7049	23.13	29
0.05	0.1	2	0.5	0.8	15	0.8211	0.7014	23.90	29
0.025	0.1	1	0.5	0.8	20	0.7901	0.7119	29.09	36
0.025	0.1	2	0.5	0.8	19	0.8180	0.7084	30.19	36
0.05	0.2	1	0.5	0.8	12	0.8176	0.7310	18.85	22
0.05	0.2	2	0.5	0.8	11	0.8554	0.7293	18.88	22
0.025	0.2	1	0.5	0.8	15	0.8211	0.7397	23.90	27
0.025	0.2	2	0.5	0.8	14	0.8515	0.7362	24.13	27
0.05	0.1	1	0.6	0.8	29	0.7882	0.7319	41.57	52
0.05	0.1	2	0.6	0.8	27	0.8138	0.7293	42.68	52
0.025	0.1	1	0.6	0.8	34	0.7959	0.7397	50.13	64
0.025	0.1	2	0.6	0.8	33	0.8146	0.7362	52.61	64
0.05	0.2	1	0.6	0.8	21	0.8131	0.7536	32.92	38
0.05	0.2	2	0.6	0.8	20	0.8383	0.7484	33.92	38
0.025	0.2	1	0.6	0.8	27	0.8138	0.7571	42.68	48
0.025	0.2	2	0.6	0.8	25	0.8372	0.7553	42.77	48
0.05	0.1	1	0.6	0.9	9	0.8851	0.8127	12.81	15
0.05	0.1	2	0.6	0.9	8	0.9166	0.8136	12.84	15
0.025	0.1	1	0.6	0.9	10	0.8978	0.8301	14.91	19
0.025	0.1	2	0.6	0.9	10	0.9149	0.8231	16.04	19
0.05	0.2	1	0.6	0.9	7	0.9069	0.8336	10.81	12
0.05	0.2	2	0.6	0.9	6	0.9394	0.8379	10.37	12
0.025	0.2	1	0.6	0.9	8	0.9166	0.8509	12.84	14
0.025	0.2	2	0.6	0.9	8	0.9327	0.8440	13.71	14
					p=3				
0.05	0.1	1	0.5	0.7	28	0.6826	0.6250	40.08	53
0.05	0.1	2	0.5	0.7	27	0.7074	0.6215	41.91	53
0.025	0.1	1	0.5	0.7	35	0.6863	0.6319	50.35	64
0.025	0.1	2	0.5	0.7	33	0.7088	0.6284	51.55	64
0.05	0.2	1	0.5	0.7	21	0.7059	0.6423	32.41	38
0.05	0.2	2	0.5	0.7	20	0.7339	0.6380	33.27	38
0.025	0.2	1	0.5	0.7	26	0.7106	0.6493	40.76	48
0.025	0.2	2	0.5	0.7	25	0.7333	0.6458	41.88	48
0.05	0.1	1	0.5	0.8	11	0.7656	0.6884	15.35	19
0.05	0.1	2	0.5	0.8	10	0.8027	0.6875	15.40	19
0.025	0.1	1	0.5	0.8	13	0.7756	0.7014	18.45	23
0.025	0.1	2	0.5	0.8	12	0.8060	0.6997	18.63	23
0.05	0.2	1	0.5	0.8	8	0.7984	0.7154	12.22	14
0.05	0.2	2	0.5	0.8	7	0.8409	0.7154	11.75	14
0.025	0.2	1	0.5	0.8	10	0.8027	0.7223	15.40	18

0.025	0.2	2	0.5	0.8	9	0.8365	0.7258	15.09	18
0.05	0.1	1	0.6	0.8	19	0.7810	0.7275	26.93	35
0.05	0.1	2	0.6	0.8	18	0.8045	0.7240	27.81	35
0.025	0.1	1	0.6	0.8	22	0.7897	0.7362	32.05	42
0.025	0.1	2	0.6	0.8	22	0.8057	0.7310	34.16	42
0.05	0.2	1	0.6	0.8	14	0.8033	0.7449	21.55	25
0.05	0.2	2	0.6	0.8	13	0.8299	0.7414	21.68	25
0.025	0.2	1	0.6	0.8	17	0.8089	0.7536	26.66	32
0.025	0.2	2	0.6	0.8	16	0.8304	0.7501	26.94	32
0.05	0.1	1	0.6	0.9	6	0.8718	0.8023	8.44	11
0.05	0.1	2	0.6	0.9	6	0.8936	0.7936	9.05	11
0.025	0.1	1	0.6	0.9	7	0.8808	0.8162	10.03	13
0.025	0.1	2	0.6	0.9	7	0.8981	0.8092	10.68	13
0.05	0.2	1	0.6	0.9	5	0.8871	0.8170	7.45	8
0.05	0.2	2	0.6	0.9	4	0.9261	0.8231	6.73	8
0.025	0.2	1	0.6	0.9	6	0.8936	0.8266	9.05	10
0.025	0.2	2	0.6	0.9	5	0.9248	0.8353	8.42	10
				p=4					
0.05	0.1	1	0.5	0.7	23	0.6773	0.6215	32.58	43
0.05	0.1	2	0.5	0.7	22	0.7022	0.6180	33.77	43
0.025	0.1	1	0.5	0.7	28	0.6831	0.6284	40.09	52
0.025	0.1	2	0.5	0.7	27	0.7034	0.6250	41.58	52
0.05	0.2	1	0.5	0.7	17	0.7010	0.6380	26.05	31
0.05	0.2	2	0.5	0.7	16	0.7297	0.6350	26.43	31
0.025	0.2	1	0.5	0.7	21	0.7061	0.6458	32.63	39
0.025	0.2	2	0.5	0.7	20	0.7292	0.6432	33.23	39
0.05	0.1	1	0.5	0.8	9	0.7581	0.6832	12.50	16
0.05	0.1	2	0.5	0.8	8	0.7970	0.6823	12.27	16
0.025	0.1	1	0.5	0.8	10	0.7743	0.6980	14.35	19
0.025	0.1	2	0.5	0.8	10	0.7962	0.6910	15.23	19
0.05	0.2	1	0.5	0.8	6	0.7991	0.7154	9.31	12
0.05	0.2	2	0.5	0.8	6	0.8272	0.7049	9.87	12
0.025	0.2	1	0.5	0.8	8	0.7970	0.7188	12.27	14
0.025	0.2	2	0.5	0.8	8	0.8194	0.7101	12.96	14
0.05	0.1	1	0.6	0.8	16	0.7759	0.7223	22.46	29
0.05	0.1	2	0.6	0.8	15	0.7997	0.7206	22.93	29
0.025	0.1	1	0.6	0.8	19	0.7828	0.7327	27.11	35
0.025	0.1	2	0.6	0.8	18	0.8024	0.7301	27.75	35
0.05	0.2	1	0.6	0.8	12	0.7962	0.7397	18.20	21
0.05	0.2	2	0.6	0.8	11	0.8232	0.7362	18.10	21
0.025	0.2	1	0.6	0.8	14	0.8048	0.7501	21.80	26
0.025	0.2	2	0.6	0.8	13	0.8271	0.7466	21.74	26
0.05	0.1	1	0.6	0.9	5	0.8650	0.7971	7.04	10
0.05	0.1	2	0.6	0.9	5	0.8867	0.7884	7.49	10
0.025	0.1	1	0.6	0.9	6	0.8718	0.8092	8.47	11
0.025	0.1	2	0.6	0.9	6	0.8891	0.8005	8.96	11
0.05	0.2	1	0.6	0.9	4	0.8834	0.8127	6.02	7
0.05	0.2	2	0.6	0.9	4	0.9048	0.8049	6.38	7
0.025	0.2	1	0.6	0.9	5	0.8867	0.8231	7.49	8
0.025	0.2	2	0.6	0.9	5	0.9036	0.8144	7.91	8