2019

Sequential Testing in Reliability and Validity Studies With Repeated Measurements per Subject

Steven B. Kim
California State University, Monterey Bay, stkim@csumb.edu

Jeffrey O. Wand
California State University, Monterey Bay, jwand@csumb.edu

Sequential Testing in Reliability and Validity Studies With Repeated Measurements per Subject

Steven B. Kim[1] & Jeffrey O. Wand[1]

[1] Mathematics and Statistics Department, California State University, Monterey Bay, Seaside, CA, USA

Correspondence: Steven B. Kim, Mathematics and Statistics Department, California State University, Monterey Bay, Seaside, CA 93955, USA. Tel: 1-831-582-3954. E-mail: stkim@csumb.edu

Received: November 19, 2018 Accepted: December 10, 2018 Online Published: December 24, 2018

doi:10.5539/ijsp.v8n1p120 URL: https://doi.org/10.5539/ijsp.v8n1p120

Abstract

In medical, health, and sports sciences, researchers desire a device with high reliability and validity. This article focuses on reliability and validity studies with n subjects and $m \geq 2$ repeated measurements per subject. High statistical power can be achieved by increasing n or m, and increasing m is often easier than increasing n in practice unless m is too high to result in systematic bias. The sequential probability ratio test (SPRT) is a useful statistical method which can conclude a null hypothesis H_0 or an alternative hypothesis H_1 with 50% of the required sample size of a non-sequential test on average. The traditional SPRT requires the likelihood function for each observed random variable, and it can be a practical burden for evaluating the likelihood ratio after each observation of a subject. Instead, m observed random variables per subject can be transformed into a test statistic which has a known sampling distribution under H_0 and under H_1. This allows us to formulate a SPRT based on a sequence of test statistics. In this article, three types of study are considered: reliability of a device, reliability of a device relative to a criterion device, and validity of a device relative to a criterion device. Using SPRT for testing the reliability of a device, for small m, results in an average sample size of about 50% of the fixed sample size for a non-sequential test. For comparing a device to criterion, the average sample size approaches to 60% approximately as m increases. The SPRT tolerates violation of normality assumption for validity study, but it does not for reliability study.

Keywords: sequential probability ratio test, reliability, validity, repeated measurements

1. Introduction

In medical, health, and sports sciences, researchers and practitioners want to use a highly valid and reliable device to conduct research or to make an important decision. Often there is a criterion devices (the standard), and researchers often test the validity and reliability of a new device against the criterion. Some researchers suggest using a correlation as a parameter of interest when two or more raters (devices) are compared (Prescott, 2018; Mokkink et al., 2010; Shrout and Fleiss, 1979). However, a correlation is difficult to interpret in the context of a research problem, and it may be controversial to set a threshold of satisfying correlation. In particular, a correlation depends on the heterogeneity of study participants (Hopkins, 2000). In this article, we model measurement error by a normal distribution, and we focus on the mean and the standard deviation (SD) of the normal model as our parameters of interest. From a statistical perspective, it is reasonable to quantify the validity by using the mean of measurement error and the reliability by the SD.

We can increase precision of parameter estimation and statistical power of hypothesis testing by increasing the sample size (i.e., the number of subjects) and/or the repetitions (i.e., the number of repeated measurements per subject). Hopkins (2000) suggested approximately 50 study participants and at least 3 trials, but the desired sample size and the number of repetitions depend on various factors. Sometimes, a too high number of repetitions may introduce systematic change in measurement (e.g., learning effect, fatigue, etc.) For some researchers, it can be difficult to recruit a large number of subjects, and most researchers would like to draw a valid conclusion in an efficient manner in terms of cost, time, and effort. For those researchers, sequential analysis can be a useful statistical method. Wald (1945) introduced the sequential probability ratio test (SPRT) which allows a researcher to terminate hypothesis testing with about 50% of the required sample size on average while preserving the desired significance level and statistical power. In some practical cases, the SPRT (which requires data monitoring after each new data point) can be a serious practical burden, and the SPRT has advanced to various forms of group sequential tests in clinical trials (Pocock, 1977; O'Brien & Fleming, 1979; Wang & Tsiatis, 1987; Jennison & Turnbull, 2000).

In some practical situations, it is more convenient to increase the number of repeated measurements per subject rather than increasing the sample size. Let n denote the sample size (i.e., the number of subjects) and m denote the number of repetitions per subject. Using these notations, there are $n \times m$ data points denoted by Y_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. It

may be challenging and exhausting to perform $n \times m$ sequential tests (after each observation for every subject), but it may be more practical to perform n sequential tests after each subject. In this article, we demonstrate the effect of increasing m for fixed n in a reliability and validity study, and we demonstrate the operating characteristics of the SPRT when it is done after each subject with m repeated measurements taken per subject. When m random variables $(Y_{i1}, \ldots, Y_{im})$ are observed from the i^{th} subject, we can apply the SPRT based on a known sampling distribution of $g(Y_{i1}, \ldots, Y_{im})$, where g: $\mathcal{R}^m \to \mathcal{R}$ (i.e., a single-value statistic which summarizes m data points). Thus, for reliability and validity studies which typically involve m repeated measurements per subject, the novelty of this article is that we apply the SPRT based on the sampling distribution of $g(Y_{i1}, \ldots, Y_{im})$ rather than the SPRT based on the distribution of each Y_{ij} (as typically done in a SPRT). We provide simulation results that this SPRT, which is based on the likelihood function of the sampling distribution of $g(Y_{i1}, \ldots, Y_{im})$, still preserves significance level and statistical power while significantly reducing the average sample size. To this end, researchers who study reliability and validity of measurement devices can terminate their studies early with a substantially fewer number of subjects by taking m repeated measurements per subject.

This manuscript is structured as follows. Section 2 includes a normal error model, some terminology used throughout this article, and a brief review of SPRT. In Section 3, we focus on hypothesis testing for the reliability of a single device. In Section 4, we discuss hypothesis testing for comparing the reliability of a new device to a criterion (i.e., comparing SDs of measurement error in two devices). In Section 5, we discuss hypothesis testing for comparing the validity of a new device to a criterion (i.e., comparing means of measurement error in two devices).

2. Assumptions, Terminology and Review of SPRT

2.1 Normal Error Model

Suppose the value of a subject is measured by a device (e.g., body temperature, body mass index, force generated by a body part, etc.). Let μ denote the true value of a subject (or simply the truth). Suppose the device measures the true value m times (repeated measurements). Let Y_j denote the observed value in the j^{th} measurement which is not exactly equal to μ. We define the j^{th} *measurement error* by $\epsilon_j = Y_j - \mu$, the difference between the observed value Y_j and the truth. A typical probability model for measurement error is a normal distribution, denoted by $\epsilon_j \sim N(\beta, \sigma^2)$. We further assume $\epsilon_1, \ldots, \epsilon_m$ are independent. There are two parameters of interest, the *mean error* β and the *standard deviation* (*SD*) σ. The normal error model is graphically represented in Figure 1. In later sections, the j^{th} measurement error from the i^{th} subject will be denoted by ϵ_{ij}, and we will assume ϵ_{ij}'s are independent from measurement to measurement and from subject to subject.

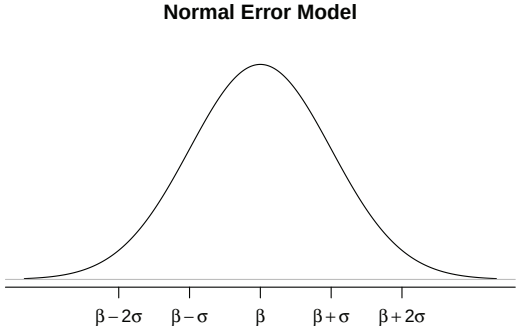

Figure 1. Graphical representation of a normal error model denoted by $N(\beta, \sigma^2)$

2.2 Terminology

We say a device is *valid* (or *unbiased*) when $\beta = 0$. When a device is invalid (i.e., $\beta \neq 0$), *overestimation* refers to $\beta > 0$, and *underestimation* refers to $\beta < 0$. We say a device is more or less *reliable* (or more or less *precise*) when σ is smaller or larger, respectively, where σ takes some positive value. For comparing two devices, device 1 relative to device 0, let β_0 and β_1 denote the mean error of device 0 and the mean error of device 1, respectively. Without loss of generality, assume that device 1 is a new device being tested and device 0 serves as a criterion (known to be standard). We say device 1 is *valid relative to device 0* when $\beta_1 - \beta_0 = 0$. Let σ_0 and σ_1 denote the SDs of device 0 and device 1, respectively. We say device 1 is *as reliable as device 0* when $\sigma_1/\sigma_0 = 1$ and *less reliable than device 0* when $\sigma_1/\sigma_0 > 1$. A study of β (or $\beta_1 - \beta_0$) is referred to as a *validity study*, and a study of σ (or σ_1/σ_0) is referred to as a *reliability study*.

2.3 Estimable Parameters

Consider a single device with the normal error assumption $\epsilon_j \sim N(\beta, \sigma^2)$. If the truth is known, a sequence of independent random variables $(\epsilon_1, \ldots, \epsilon_m)$ is observable, and unbiased estimators of β and σ^2 are $\bar{\epsilon} = \frac{1}{m}\sum_{j=1}^{m} \epsilon_j$ and $S_\epsilon^2 = \frac{1}{m-1}\sum_{j=1}^{m}(\epsilon_j - \bar{\epsilon})^2$, respectively (Hogg and Tanis, 1997). If the truth is unknown, $(\epsilon_1, \ldots, \epsilon_m)$ is not observable, but $(Y_1, \ldots, Y_m)$ is observable instead. With $Y_j = \mu + \epsilon_j$ for $j = 1, \ldots, m$, we cannot estimate β, but we can estimate σ^2 by $S_Y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \bar{Y})^2$, where $\bar{Y} = \frac{1}{m}\sum_{j=1}^{m} Y_j$ is the sample mean which estimates $\mu + \beta$. In most practical cases, the truth is unknown, so only σ can be estimated. For comparing two devices with normal error models $N(\beta_0, \sigma_0^2)$ for device 0 and $N(\beta_1, \sigma_1^2)$ for device 1, we can estimate $\beta_1 - \beta_0$ and σ_1/σ_0 even when the truth is unknown.

2.4 Practical Research Questions for Studying a Single Device

In most practical situations, the truth is unknown, so it is impossible to answer the research question "Is this device valid?" without a criterion. In other words, in absence of knowing the true value of a subject, we cannot conduct a validity study of a single device. Since there is no device with perfect reliability (i.e., $\sigma > 0$), it is challenging to address whether a device is reliable or not without some threshold. Therefore, a practical research question is "Do we have statistical evidence to reject $\sigma = \sigma_0$ in favor of $\sigma > \sigma_0$?" where σ_0 is a maximal acceptable SD. For example, a 100-meter sprint world record is rounded to two decimal places in seconds. In order to record the truth 9.58 seconds as 9.58 by a valid device with a probability 0.9973 (within 3 SDs from the truth by the empirical rule of a normal model), we want to find σ_0 such that $9.58 - 3\sigma_0 = 9.575$ or $9.58 + 3\sigma_0 = 9.584\bar{9}$, so $\sigma_0 = 0.001\bar{6}$.

2.5 Review of SPRT

Wald (1945) introduced the sequential probability ratio test (SPRT) which allows interim hypothesis testing after each observation. In the framework of SPRT, the sample size is a random variable, and we use an uppercase letter N to denote the random sample size. The general procedure of SPRT is as follows. Let L_{0i} and L_{1i} denote the likelihood after the i^{th} subject under a simple null hypothesis H_0 and a simple alternative hypothesis H_1, respectively. Let α and $1 - \zeta$ denote the fixed significance level and desired statistical power, respectively. Let $\Lambda_i = L_{1i}/L_{0i}$ be the likelihood ratio which serves as the test statistic for SPRT, and one of the following three decisions is made after observing each i^{th} subject:

- Case 1: Terminate the study by concluding H_0 if $\Lambda_i \leq \frac{\zeta}{1-\alpha}$.
- Case 2: Terminate the study by concluding H_1 if $\Lambda_i \geq \frac{1-\zeta}{\alpha}$.
- Case 3: Do not make any conclusion and continue the study if $\frac{\zeta}{1-\alpha} \leq \Lambda_i \leq \frac{1-\zeta}{\alpha}$.

A SPRT is guaranteed to make a conclusion with a finite sample size (Wald, 1945). As compared with the fixed sample size under the most powerful test introduced by Neyman (1933), the SPRT often results in a saving of about 50% in the sample size on average (Wald, 1945).

3. Reliability Study of a Single Device With *n* Subjects and *m* Repetitions

3.1 Formulation of Hypothesis Testing

In a reliability study for a single device (no criterion device), suppose n subjects are recruited and m repeated measurements are taken per subject. In most practical situations, the truth varies from subject to subject (e.g., body mass index). Let μ_i denote the unknown fixed truth of the i^{th} subject for $i = 1, \ldots, n$. We do not make any assumption about the distribution of μ_i because it is not needed in our discussion. Let Y_{ij} denote the observed value of the j^{th} measurement from the i^{th} subject, and the $(i, j)^{\text{th}}$ measurement error is denoted by $\epsilon_{ij} = Y_{ij} - \mu_i$. As discussed in Section 2.1, we assume $\epsilon_{ij} \sim N(\beta, \sigma^2)$ and independence among ϵ_{ij}'s for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. The null hypothesis is H_0: $\sigma = \sigma_0$, and the alternative hypothesis is H_1: $\sigma > \sigma_0$, where σ_0 is a maximal SD under practical considerations.

3.2 Exact Sampling Distribution

Let $(Y_{i1}, \ldots, Y_{im})$ be m random variables observed from the i^{th} subject. Let $\bar{Y}_i = \frac{1}{m}\sum_{j=1}^{m} Y_{ij}$ be the sample mean for the i^{th} subject, and let $S_i^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_{ij} - \bar{Y}_i)^2$ be the sample variance which is an unbiased estimator for σ^2 based on the m observations made from the ith subject. We have the exact sampling distribution

$$W_i = \frac{(m-1)S_i^2}{\sigma^2} \sim \chi_{m-1}^2 ,$$

where χ_{m-1}^2 denotes the chi-square distribution with $m - 1$ degrees of freedom (Hogg and Tanis, 1997). Since $S_1^2, \ldots, S_n^2$

are independent random variables,

$$W = \frac{n(m-1)S_{\cdot}^{2}}{\sigma^2} \sim \chi^2_{n(m-1)}\,,$$

where $S_{\cdot}^{2} = \frac{1}{n}\sum_{i=1}^{n} S_i^2$ is an unbiased estimator for σ^2 when all data points from n subjects are accumulated. For the purpose of hypothesis testing with H_0: $\sigma = \sigma_0$, we replace the unknown parameter σ by the null value σ_0.

The effect of increasing m for fixed n is as follows. Given the true value of σ, at significance level α, the statistical power shall depend on the product $n(m-1)$. If $m = 2^j + 1$ for $j = 0, 1, 2, \ldots$, then for $m^* = 2^{j+1} + 1$

$$W = \frac{\frac{n}{2}(m^*-1)S_{\cdot}^{2}}{\sigma^2} \sim \chi^2_{\frac{n}{2}(m^*-1)}\,.$$

This follows from the fact that

$$\begin{aligned} n(m-1) &= n(2^j) \\ &= \frac{n}{2}(2)(2^j) \\ &= \frac{n}{2}(2^{j+1}+1-1) \\ &= \frac{n}{2}(m^*-1)\,. \end{aligned}$$

In other words, if we increase the number of repeated measurements per subject from m to m^* for any j (e.g., from 2 to 3, from 3 to 5, from 5 to 9, and so on), we can maintain the statistical power with one half of n.

3.3 Power Analysis in Non-Sequential Test for σ

If a researcher specifies H_0: $\sigma = \sigma_0$ under practical considerations, it is reasonable to consider a one-sided alternative hypothesis H_1: $\sigma > \sigma_0$. For illustration purposes, consider the significance level $\alpha = 0.05$, the null value $\sigma_0 = 0.05$, and the alternative value $\sigma_1 = 0.06$. Let $1-\zeta$ denote the statistical power. Table 1 provides required n for given m and desired $1-\zeta$ for a non-sequential test. For $\alpha = 0.05$ and $1-\zeta$, there are various designs (m, n) such as (2, 164), (3, 82), and (5, 41) to list a few.

Table 1. Required sample size n for given $1-\zeta$ and m in a non-sequential test at level $\alpha = 0.05$ with parameter values $\sigma_0 = 0.05$, and $\sigma_1 = 0.06$

$1-\zeta$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$	$m=7$	$m=8$	$m=9$	$m=10$
0.70	68	34	23	17	14	12	10	9	8
0.80	91	46	31	23	19	16	13	12	11
0.90	128	64	43	32	26	22	19	16	15
0.95	164	82	55	41	33	28	24	21	19
0.99	244	122	82	61	49	41	35	31	28

3.4 SPRT for σ

Under the assumption of H_0: $\sigma = \sigma_0$, we have the exact sampling distribution

$$T_i = \frac{(m-1)S_i^2}{\sigma_0^2} \sim \chi^2_{m-1}\,.$$

Since $T_1 = t_1, T_2 = t_2, \ldots, T_i = t_i$ are independent observations, the likelihood under H_0 is given by

$$L_{0i} = \prod_{h=1}^{i} \frac{1}{2^{\frac{m-1}{2}}\,\Gamma\left(\frac{m-1}{2}\right)}\, t_h^{\frac{m-1}{2}-1}\, e^{-\frac{t_h}{2}}\,.$$

Under the assumption of H_1: $\sigma = \sigma_1$, the exact sampling distribution is $T_i \sim \text{Gamma}\left(\frac{m-1}{2}, \frac{\sigma_0^2}{2\sigma_1^2}\right)$, so the likelihood under H_1 is given by

$$L_{1i} = \prod_{h=1}^{i} \frac{1}{2^{\frac{m-1}{2}}\,\Gamma\left(\frac{m-1}{2}\right)} \left(\frac{\sigma_0}{\sigma_1}\right)^{2(m-1)} t_h^{\frac{m-1}{2}-1}\, e^{-\frac{1}{2}\left(\frac{\sigma_0}{\sigma_1}\right)^2 t_h}\,.$$

The i^{th} test statistic for SPRT is

$$\Lambda_i = \frac{L_{1i}}{L_{0i}} = \left(\frac{\sigma_0}{\sigma_1}\right)^{2i(m-1)} e^{\frac{1}{2}\left[1-\left(\frac{\sigma_0}{\sigma_1}\right)^2\right]\sum_{h=1}^{i} t_h}$$

which is in terms of the sum $\sum_{h=1}^{i} T_h$.

To demonstrate operating characteristics of SPRT, assume $\alpha = 0.05$, $\sigma_0 = 0.05$, $\sigma_1 = 0.06$, and $1 - \zeta = 0.95$ (desired power). Using a simulation of 10,000 replicates, Tables 2 and 3 are generated. Tables 2 represents the probability of rejecting H_0, and Table 3 represents the average sample size $E(N)$. For $m \geq 2$, the probability of rejecting H_0 is slightly under $\alpha = 0.05$ when H_0: $\sigma = 0.05$ is true, and it is slightly above $1 - \zeta = 0.95$ when H_1: $\sigma = 0.06$ is true. When H_0 is true, comparing to required n in a non-sequential test for given m, the average sample size is about 57–59% for $m = 2$ and $m = 3$, and it reaches up to about 65% as m increases. When H_1 is true, the average sample size is about 49–50% for $m = 2$ and $m = 3$, and it reaches up to about 60% as m increases. This tendency is graphically demonstrated in Figure 2. The average sample size $E(N)$ is significant even when σ is slightly below $\sigma_0 = 0.05$ and when σ is slightly above as demonstrated in Table 3. The R code for the simulation study is given in Appendix 1.

We can see that $E(N)$ is nearly halved when m is increased in certain ways. If the number of repeated measurements has the form $m = 2^j + 1$ for $j = 0, 1, 2, \ldots$, the average sample size $E(N)$ seems to be halved if m is increased to $m^* = 2^{j+1} + 1$. A researcher who plans a non-sequential test which requires $n = 164$ with $m = 2$ to detect a practical significant difference between $\sigma_0 = 0.05$ and $\sigma_1 = 0.06$ with $\alpha = 0.05$ and $1 - \beta = 0.95$; however, the SPRT with $m = 3$ would result in an expected sample size which is about one quarter of n.

Table 2. Probability of rejecting H_0 in SPRT designed for $\sigma_0 = 0.05$, $\sigma_1 = 0.06$, $\alpha = 0.05$ and $1 - \zeta = 0.95$

σ	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$	$m = 8$	$m = 9$	$m = 10$
0.03	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.04	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.05	0.036	0.034	0.031	0.031	0.031	0.029	0.031	0.030	0.028
0.06	0.951	0.956	0.955	0.959	0.960	0.962	0.963	0.962	0.963
0.07	1.000	0.999	1.000	1.000	0.999	1.000	1.000	1.000	1.000
0.08	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 3. Average sample size $E(N)$ in SPRT designed for $\sigma_0 = 0.05$, $\sigma_1 = 0.06$, $\alpha = 0.05$ and $1 - \zeta = 0.95$

σ	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$	$m = 8$	$m = 9$	$m = 10$
0.03	23.8	12.1	8.3	6.3	5.2	4.4	3.8	3.4	3.1
0.04	35.6	18.2	12.4	9.4	7.7	6.5	5.6	5.0	4.5
0.05	93.9	48.2	32.8	25.0	20.4	17.0	14.8	13.2	11.8
0.06	80.5	41.4	28.0	21.3	17.4	14.7	12.8	11.3	10.2
0.07	29.5	15.0	10.3	7.9	6.5	5.5	4.8	4.3	3.9
0.08	17.1	8.9	6.2	4.8	3.9	3.4	3.0	2.7	2.4

4. Testing Reliability of One Device Relative to Criterion Device

4.1 Formulation of Hypothesis Testing

Suppose reliability of a new device (called device 1) is tested against a criterion device (called device 0) based on n subjects with m repeated measurements taken per subject. Let μ_i denote the truth of the i^{th} subject for $i = 1, \ldots, n$. Assume μ_i varies from subject to subject, and the value of μ_i is fixed and unknown. Let ϵ_{ijk} denote the measurement error in the $(i, j)^{\text{th}}$ observation by device k for $i = 1, \ldots, n$, $j = 1, \ldots, m$, and $k = 0, 1$. Assume $\epsilon_{ijk} \sim N(\beta_k, \sigma_k^2)$ and independence among all ϵ_{ijk}'s. A researcher observes the value of $Y_{ijk} = \mu_i + \epsilon_{ijk}$. It is impossible to observe μ_i and ϵ_{ijk} separately, but we can estimate σ_k^2 for $k = 0, 1$. In this section, the parameter of interest is $\tau = \sigma_1/\sigma_0$ which quantifies the *reliability of device 1 relative to device 0*. We say *device 1 is as reliable as device 0* if $\tau = 1$, and we say *device* 1 *is less reliable than device* 0 if $\tau > 1$. In terms of percentage, the SD of device 1 is $(\tau - 1)100\%$ greater than the SD of device 0. The null hypothesis is H_0: $\tau = \tau_0$, and the alternative hypothesis is H_1: $\tau = \tau_1$.

4.2 Exact Sampling Distribution

Let $(Y_{i10}, \ldots, Y_{im0})$ be m random variables generated by device 0 and $(Y_{i11}, \ldots, Y_{im1})$ be m random variables generated by device 1 when the two devices measured the i^{th} subject for $i = 1, \ldots, n$. Let $\bar{Y}_{ik} = \frac{1}{m}\sum_{j=1}^{m} Y_{ijk}$ be the sample mean for

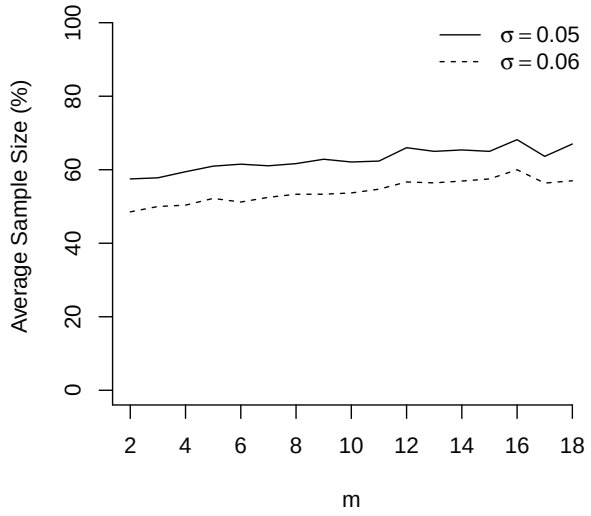

Figure 2. Average sample size of SPRT (%) with respect to m (repetitions per subject)

the i^{th} subject measured by device k. From the i^{th} subject, an unbiased estimator for σ_k^2 is $S_{ik}^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_{ijk} - \bar{Y}_{ik})^2$ for $k = 0, 1$, and we have the exact sampling distribution

$$W_i = \left(\frac{S_{i1}^2}{S_{i0}^2}\right)\left(\frac{\sigma_0^2}{\sigma_1^2}\right) = \left(\frac{S_{i1}^2}{S_{i0}^2}\right)\left(\frac{1}{\tau^2}\right) \sim \mathcal{F}_{m-1,m-1}\,,$$

where $\mathcal{F}_{m-1,m-1}$ denotes the F distribution with $m-1$ numerator degrees of freedom and $m-1$ denominator degrees of freedom. Further note that

$$W = \left(\frac{S_{\cdot 1}^2}{S_{\cdot 0}^2}\right)\left(\frac{1}{\tau^2}\right) \sim \mathcal{F}_{n(m-1),n(m-1)}\,,$$

where $S_{\cdot 1}^2 = \frac{1}{n}\sum_{i=1}^{n} S_{ik}^2$ is an unbiased estimator for σ_k^2 for $k = 0, 1$ when all data points from n subjects are combined.

4.3 Power Analysis in Non-Sequential Test for τ

For illustration purposes, consider the significance level $\alpha = 0.05$, the null value $\tau_0 = 1$, and the alternative value $\tau_1 = 1.2$. Let $1-\zeta$ denote statistical power. Table 4 presents required n for given $1-\zeta$ for $n \geq 5$ and $m = 2, 3, \ldots, 10$. There are various designs (m, n) for achieving $1-\zeta = 0.95$ such as (2, 327), (3, 164), and (5, 82). As seen in Section 3.3, a similar pattern of halving n is observed.

Table 4. Required sample size n for given $1-\zeta$ and m in a non-sequential test at level $\alpha = 0.05$ with parameter values $\tau_0 = 1$, and $\tau_1 = 1.2$

$1-\zeta$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$	$m=7$	$m=8$	$m=9$	$m=10$
0.70	143	72	48	36	29	24	21	18	16
0.80	187	94	63	47	38	32	27	24	21
0.90	259	130	87	65	52	44	37	33	29
0.95	327	164	109	82	66	55	47	41	37
0.99	476	238	159	119	96	80	68	60	53

4.4 SPRT for τ

Let

$$T_i = \left(\frac{S_{i1}^2}{S_{i0}^2}\right)\left(\frac{1}{\tau_0^2}\right) = \frac{S_{i1}^2}{S_{i0}^2}$$

be the test statistic based on m observations made on the i^{th} subject, where a researcher often sets $\tau_0 = 1$ for H_0. Under the assumption of H_0: $\tau = \tau_0$, we have the exact sampling distribution $T_i \sim \mathcal{F}_{m-1,m-1}$. Since n test statistics are independent, the likelihood under H_0 is given by

$$L_{0i} = \prod_{h=1}^{i} \frac{\Gamma(m-1)}{\left(\Gamma\left(\frac{m-1}{2}\right)\right)^2} t_h^{\frac{m-1}{2}-1} (1+t_h)^{-(m-1)} .$$

Under the assumption of H_1: $\tau = \tau_1$,

$$T_i^* = \left(\frac{S_{i1}^2}{S_{i0}^2}\right)\left(\frac{1}{\tau_1^2}\right) \sim \mathcal{F}_{m-1,m-1} .$$

By letting $c = (\tau_1/\tau_0)^2$, and using the Jacobian transformation, the PDF of $T_i = cT_i^*$ is written as

$$f(t_i) = \frac{\Gamma(m-1)}{\left(\Gamma\left(\frac{m-1}{2}\right)\right)^2} t_i^{\frac{m-1}{2}-1} \left(1+\frac{t_i}{c}\right)^{-(m-1)} \left(\frac{1}{c}\right)^{\frac{m-1}{2}}$$

which is known as the generalized F distribution, denoted by G3F(α, β, λ) with $\alpha = \beta = (m-1)/2$ and $\lambda = 1/c$ (Pham-Gia and Duong, 1989). Therefore, the likelihood under H_1 is given by

$$L_{1i} = \prod_{h=1}^{i} \frac{\Gamma(m-1)}{\left(\Gamma\left(\frac{m-1}{2}\right)\right)^2} t_h^{\frac{m-1}{2}-1} \left(1+\frac{t_h}{c}\right)^{-(m-1)} \left(\frac{1}{c}\right)^{\frac{m-1}{2}} ,$$

and the i^{th} test statistic for SPRT is

$$\Lambda_i = \frac{L_{1i}}{L_{0i}} = c^{-\frac{i(m-1)}{2}} \left(\prod_{h=1}^{i} z_h\right)^{m-1} ,$$

where $z_i = \frac{1+t_i}{1+t_i/c}$. Assuming $\tau_0 = 1$, an alternative form of the i^{th} test statistic is

$$\Lambda_i = \left(\frac{1}{\tau_1}\right)^{i(m-1)} \prod_{h=1}^{i} \left(\frac{s_{h0}^2 + s_{h1}^2}{s_{h0}^2 + (s_{h1}/\tau_1)^2}\right)^{m-1} .$$

The operating characteristics of the SPRT for τ were studied using a simulation of 10,000 replicates with $\alpha = 0.05$, $\tau_0 = 1$, $\tau_1 = 1.2$, and $1 - \zeta = 0.95$. Tables 5 and 6 represent the probability of rejecting H_0 and $E(N)$, respectively. For small m, particularly $m = 2$, the SPRT based on the exact sampling distribution of T_i did not follow the usual characteristics of Wald's SPRT (which often results in about 50% average sample size while preserving α and $1 - \zeta$). When $m = 2$ and H_0: $\tau = 0.1$ is true, H_0 is rejected with a probability 0.073 which is slightly greater than the fixed $\alpha = 0.05$. When $m = 2$ and $\tau = 1.2$ is true, the resulting statistical power is 0.915 which is lower than the fixed $1 - \zeta = 0.95$. Furthermore, when $m = 2$ and τ is near 1 and 1.2, $E(N)$ is greater than $n = 327$ (the sample size required for a non-sequential test) which defeats the purpose of SPRT. On the other hand, for $m \geq 3$, the Type I error probability is at most $\alpha = 0.05$, and statistical power is at least $1 - \zeta = 0.95$. For large m, as shown in Figure 3, the average sample size is close to 60% when $\tau = 1$ and $\tau = 1.2$, but it does not seem to go below 60% as m increases. (We simulated up to $m = 19$.) The R code for the simulation study is given in Appendix 2.

5. Testing Validity of One Device Relative to Criterion Device

5.1 Formulation of Hypothesis Testing

Suppose the validity of a new device (device 1) is tested against a criterion device (device 0) with n subjects and m repeated measurements per subject. Let μ_{ij} denote the truth of the i^{th} subject in the j^{th} measurement. In other words, the truth may vary from subject to subject and from trial to trial within subject. Let ϵ_{ijk} denote the measurement error in the $(i, j)^{\text{th}}$ measurement by device k. Assume $\epsilon_{ijk} \sim N(\beta_k, \sigma_k^2)$ and independence among all ϵ_{ijk}'s. A researcher observes the value of $Y_{ijk} = \mu_{ij} + \epsilon_{ijk}$. Even though it is impossible to observe μ_{ij} and ϵ_{ijk} separately, we can still estimate $\beta_1 - \beta_0$. (We cannot

Table 5. Probability of rejecting H_0 in SPRT designed for $\tau_0 = 1$, $\tau_1 = 1.2$, $\alpha = 0.05$ and $1 - \zeta = 0.95$

τ	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$	$m = 8$	$m = 9$	$m = 10$
0.8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.9	0.003	0.002	0.001	0.001	0.002	0.001	0.002	0.001	0.001
1.0	0.073	0.048	0.044	0.040	0.041	0.042	0.040	0.038	0.038
1.1	0.502	0.515	0.527	0.536	0.527	0.533	0.533	0.533	0.538
1.2	0.915	0.953	0.955	0.958	0.962	0.960	0.960	0.963	0.961
1.3	0.991	0.997	0.997	0.997	0.997	0.997	0.998	0.997	0.997
1.4	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 6. Average sample size $E(N)$ in SPRT designed for $\tau_0 = 1$, $\tau_1 = 1.2$, $\alpha = 0.05$ and $1 - \zeta = 0.95$

τ	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$	$m = 8$	$m = 9$	$m = 10$
0.8	110.1	40.8	24.4	17.5	13.6	11.2	9.6	8.3	7.4
0.9	315.6	64.3	38.9	27.6	21.4	17.6	15.0	12.9	11.5
1.0	1110.8	141.6	76.7	54.6	42.1	34.6	29.0	25.8	22.4
1.1	1674.0	295.4	133.9	91.0	72.1	59.2	50.3	44.0	38.6
1.2	1277.3	144.0	77.5	53.9	41.9	34.4	29.7	25.5	22.7
1.3	589.6	73.8	44.3	31.3	24.5	19.9	17.1	15.0	13.1
1.4	242.5	51.8	31.6	22.2	17.4	14.2	12.0	10.4	9.3

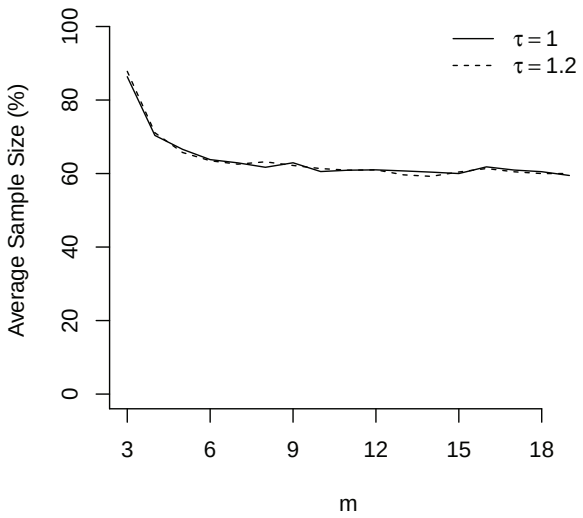

Figure 3. Average sample size of SPRT (%) with respect to m (repetitions per subject)

estimate β_1 and β_0 separately unless the mean error β_0 is known for device 0 or β_0 is assumed to be zero or a specific value.) In this section, the parameter of interest is $\theta = \beta_1 - \beta_0$ which quantifies the average deviation of measurement by device 1 from measurement by device 0. We say *device 1 is valid relative to device 0* if $\theta = 0$, and we say *device 1 is not valid relative to device 0* if $\theta \neq 0$. In particular, $\theta > 0$ ($\theta < 0$) means overestimation (underestimation) by device 1 when compared to device 0 on average. The null hypothesis is H_0: $\theta = 0$, and the alternative hypothesis is H_1: $\theta > 0$ or $\theta < 0$ depending on researcher's knowledge and/or purpose.

5.2 Exact Sampling Distribution

Let $D_{ij} = Y_{ij1} - Y_{ij0}$ be the difference in the observed values when device 1 is compared to device 0. Since both Y_{ij0} and Y_{ij1} have the same target μ_{ij}, the difference is $D_{ij} = \epsilon_{ij1} - \epsilon_{ij0} \sim N(\theta, \delta^2)$, where $\theta = \beta_1 - \beta_0$ and $\delta^2 = \sigma_0^2 + \sigma_1^2$. Let $\bar{D}_i = \frac{1}{m}\sum_{j=1}^{m} D_{ij}$ which is an unbiased estimator for θ based on m data points obtained from the i^{th} subject. Let

$S_i^2 = \frac{1}{m-1}\sum_{j=1}^{m}(D_{ij} - \bar{D}_i)^2$, which is an unbiased estimator for δ^2 based on m data points obtained from the i^{th} subject. We have the exact sampling distribution

$$T_i = \frac{\bar{D}_i - \theta}{S_i/\sqrt{m}} \sim \mathcal{T}_{m-1}\,,$$

where $\mathcal{T}_{m-1}$ denotes the T distribution with $m-1$ degrees of freedom. Further note that

$$T = \frac{\bar{D}_\cdot - \theta}{S_\cdot/\sqrt{m}} \sim \mathcal{T}_{n-1}\,,$$

where $\bar{D}_\cdot = \frac{1}{n}\sum_{i=1}^{n}\bar{D}_i$ and $S_\cdot^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{D}_i - \bar{D}_\cdot)^2$.

5.3 Power Analysis in Non-Sequential Test for θ

For illustration purposes, consider the significance level $\alpha = 0.05$, the null value $\theta_0 = 0$, the alternative value $\theta_1 = 0.1$, and the SDs $\sigma_0 = 0.16$ and $\sigma_1 = 0.20$ for devices 0 and 1, respectively, so that $\delta^2 = 0.16^2 + 0.2^2$. Let $1-\zeta$ denote statistical power. Table 7 provides required sample size n for given $1-\zeta$ and m for a non-sequential test (for a one-sided test and a two-sided test).

Table 7. Required sample size n for given $1-\zeta$ and m in a non-sequential test at level $\alpha = 0.05$ with parameter values $\theta_0 = 0$, $\theta_1 = 0.1$, $\sigma_0 = 0.16$, and $\sigma_1 = 0.2$

$1-\zeta$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$	$m=7$	$m=8$	$m=9$	$m=10$
0.70	17	12	10	8	7	6	6	6	5
0.80	22	15	12	10	9	8	7	7	6
0.90	30	21	16	13	11	10	9	8	8
0.95	37	26	20	16	14	12	11	10	9
0.99	54	36	28	23	19	17	15	14	12

5.4 SPRT for θ

Let

$$T_i = \frac{\bar{D}_i - \theta_0}{S_i/\sqrt{m}}$$

be the test statistic based on m random variables $(D_{i1}, \ldots, D_{im})$ observed from the i^{th} subject. Under the assumption of H_0: $\theta = \theta_0$, we have the exact sampling distribution $T_i \sim \mathcal{T}_{m-1}$. Since $T_1, T_2, \ldots, T_i$ are independent random variables, the likelihood under H_0 is given by

$$L_{0i} = \prod_{h=1}^{i} \frac{\Gamma\left(\frac{m}{2}\right)}{\sqrt{(m-1)\pi}\,\Gamma\left(\frac{m-1}{2}\right)}\left(1 + \frac{t_h^2}{m-1}\right)^{-\frac{m}{2}}.$$

Under the assumption of H_1: $\theta = \theta_1$, the exact sampling distribution is $T_i \sim \mathcal{T}_{m-1,c}$, the non-central T distribution with $m-1$ degrees of freedom and non-centrality parameter

$$c = \frac{\sqrt{m}(\theta_1 - \theta_0)}{\sqrt{\sigma_0^2 + \sigma_1^2}}\,.$$

The likelihood under H_1 is given by $L_{1i} = \prod_{h=1}^{i} f(t_h)$, where f is the PDF of $\mathcal{T}_{m-1,c}$ (Lenth, 1989; Johnson et al., 1995). Then the i^{th} test statistic for SPRT is $\Lambda_i = L_{1i}/L_{0i}$.

For a simulation study of 10,000 replicates, we let $\alpha = 0.05$, $\theta_0 = 0$, $\theta_1 = 0.1$, $\sigma_0 = 0.16$ for device 0, $\sigma_1 = 0.2$ for device 1, and $1-\zeta = 0.95$. Tables 8 and 9 represent the probability of rejecting H_0 and $E(N)$, respectively. For $m \geq 2$, the probability of Type I error is at most $\alpha = 0.05$ when $\theta = \theta_0 = 0$ and statistical power is at least $1-\zeta = 0.95$ when $\theta = \theta_1 = 0.1$. When H_0: $\theta = \theta_0 = 0$ is true, the average sample size is about 66% for $m = 2$, and it approaches to 55% as m increases. When H_1: $\theta = \theta_1 = 0.1$ is true, it is about 69% for $m = 2$ and approaches to 58% as m increases. When the true value of θ is at the exact midpoint of the null value 0 and the alternative value 0.1, the average sample size is greater than the fixed sample size n for a non-sequential test. When the true value of θ is greater than 0.1, the average sample size is significantly lower than 50% of n (e.g., as low as 25 – 28% when $\theta = 0.2$)as shown in Figure 4. The R code for this simulation study is given in Appendix 3.

Table 8. Probability of rejecting H_0 in SPRT designed for $\theta_0 = 0$, $\theta_1 = 0.1$, $\sigma_0 = 0.16$, $\sigma_1 = 0.20$, $\alpha = 0.05$ and $1 - \zeta = 0.95$

θ	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$	$m = 8$	$m = 9$	$m = 10$
0.00	0.044	0.038	0.035	0.033	0.036	0.029	0.030	0.024	0.028
0.05	0.516	0.498	0.507	0.503	0.501	0.495	0.505	0.504	0.508
0.10	0.962	0.964	0.966	0.968	0.968	0.970	0.973	0.973	0.974
0.15	0.998	0.998	0.998	0.999	1.000	0.999	0.999	0.999	1.000
0.20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 9. Average sample size $E(N)$ in SPRT designed for $\theta_0 = 0$, $\theta_1 = 0.1$, $\sigma_0 = 0.16$, $\sigma_1 = 0.20$, $\alpha = 0.05$ and $1 - \zeta = 0.95$

θ	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$	$m = 8$	$m = 9$	$m = 10$
0.00	24.4	15.8	11.8	9.5	7.9	6.9	6.1	5.5	5.0
0.05	42.9	27.7	21.0	17.1	14.6	12.5	11.1	10.2	9.3
0.10	25.4	16.1	12.2	9.8	8.2	7.1	6.3	5.7	5.2
0.15	14.5	9.4	7.0	5.6	4.7	4.1	3.7	3.3	3.1
0.20	10.4	6.8	5.1	4.2	3.5	3.1	2.7	2.5	2.3

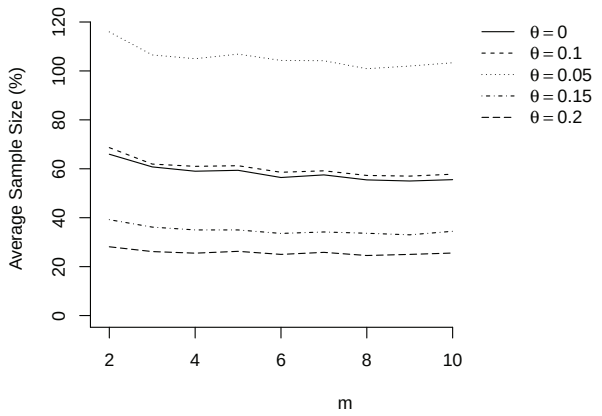

Figure 4. Average sample size SPRT (%) with respect to m (repetitions per subject)

6. Impact of Violating Normality Assumption

In practice, the true distribution of measurement error may deviate from a normal distribution. To investigate the impact of violating the normality assumption on the probability of rejecting H_0 and the average sample size $E(N)$, measurement error is simulated from a Gumbel distribution and mixed exponential distributions with a gradual increase of skewness (see Appendix 4 for more detail). When studying relative validity (Section 5.4), values of model parameters are chosen to satisfy $\sigma_0 = 0.16$, $\sigma_1 = 0.20$, and zero mean for the null case, $\theta = 0$, for the Gumbel and mixed exponential distributions. In a mixed exponential distribution, the parameter values are gradually altered to generate different skewness. For the alternative case $\theta = 0.1$, a distribution with zero mean is shifted by 0.1 for device 1. As shown in Tables 10 and 11 of Appendix 4, the Type I error probability is below the fixed $\alpha = 0.05$, the statistical power is above the fixed $1 - \zeta = 0.95$, and the average sample size $E(N)$ is close or slightly lower than the case of normality. However, when studying reliability (Section 3.4) and relative reliability (Section 4.4), the Type I error probability is about 2 – 3 times $\alpha = 0.05$, and the statistical power is below $1 - \zeta$. Researchers conducting reliability studies should be aware of the impact of violating the normality assumption, and reasons and potential remedies remain in our future study.

7. Discussion

As devices for medical and sports science are being more accurate and precise, detecting superiority or inferiority of a new device requires a large number of subjects. When it is appropriate under practical considerations, it would be cost

effective to increase m, the number of repetitions, and to implement SPRT based on the exact sampling distribution of a test statistic calculated after each subject. The SPRT for σ does not require any assumption about β, and the average sample size is about 57–58% for small m under H_0 and about 49–50% for small m under H_1. However, the SPRT for σ cannot be recommended with $m = 2$ because it results in Type I error rate greater than α and statistical power lower than $1 - \zeta$ with an average sample size being greater than the sample size required for a non-sequential test. The SPRT for τ does not require any assumption about β_0 and β_1, and the average sample size approaches to about 60% under H_0 or H_1 as m increases. The SPRT for θ does not require any assumption about σ_0 and σ_1, and the average sample size approaches to 55–60% under H_0 or H_1 as m increases. For any SPRT considered in this article, the average sample size becomes significantly lower with a slight deviation from the null value or the alternative value.

For the SPRT for θ, under the model assumptions, the truth μ_{ij} may vary from subject to subject and from trial to trial. If it is the case for testing σ or τ, the variation of μ_{ij} within subject (between trials) becomes a part of the reliability measure. For example, if a device measures walking speed of a subject m times and the variance of μ_{ij} is ν^2 within subject, the SPRT for τ would test for $\sqrt{(\sigma_1^2 + \nu^2)/(\sigma_0^2 + \nu^2)}$ which is always smaller than σ_1/σ_0. Therefore, the reliability of truth within subject becomes critical when we compare reliability of two devices using the SPRT.

Wald (1945) approximated the average savings of SPRT would be about 50% in the case when the random variable is normally distributed. In particular, he analytically approximated the expected value of the sample size under both the null and alternative hypothesis in order to approximate average savings. We are especially interested in why the SPRT based on the exact sampling distribution results in the varying savings of sample size in the reliability and validity studies with m repeated measurements. In the future, we would also like to extend Wald's calculation to the sampling distributions considered in this article (e.g., χ^2, non-central T, or F distribution).

There are other parameters of interest when two or more devices are compared such as intraclass correlation and Cronbach's alpha (Cronbach, 1951; Shrout, 1979). Kistner and Muller (2004) provided the exact sampling distributions under normality assumption and general covariance structure. Therefore, we may be able to use the sampling distributions of their estimators to study the operating characteristics of SPRT. Jin et al. (2013) derived group sequential testing with two groups (referred to as two-stage design) for interclass reliability, and they showed the average sample size is about 64–86% depending on study design parameters. A potential direction of future studies is to compare the SPRT for these parameters to the two-stage design.

References

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334. https://doi.org/10.1007/BF02310555

Hogg, R. V., & Tanis, E. A. (1997). Probability and statistical inference. Upper Saddle River, NJ: Prentice Hall.

Hopkins, W. G. (2000). Measures of reliability in sports medicine and science. *Sports Medicine, 30*(1), 1-15. https://doi.org/10.2165/00007256-200030010-00001

Jennison, C., & Turnbull, B. W. (2000). Group sequential methods with applications to clinical trials. Boca Raton: Chapman & Hall/CRC, NJ: Prentice Hall.

Jin, M., Liu, A., Chen, Z., & Li, Z. (2013). Sequential testing of measurement errors in inter-rater reliability studies. *Statistica Sinica, 23*, 1743-1759. https://doi.org/10.5705/ss.2012.036s

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). Continuous univariate distributions: volume 2. Wiley, New York.

Kistner, E. O., & Muller, K. E. (2004). Exact distributions of intraclass correlation and Cronbach's alpha with Gaussian data and general covariance. *Psychometrika, 69*(3), 459-474. https://doi.org/10.1007/BF02295646

Lenth, R. V. (1989). Algorithm AS 243 - Cumulative distribution function of the non-central t distribution. *Applied Statistics, 38*, 185-189. https://doi.org/10.2307/2347693

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., ... & De Vet, H. C. W. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of Life Research, 19(4)*, 539-549. https://doi.org/10.1007/s11136-010-9606-8

Neyman, J., & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A, 231*.

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics, 35(3)*, 549-556. https://doi.org/10.2307/2530245

Pham-Gia, T., & Duong, Q. P. (1989). The generalized beta- and F-distributions in statistical modeling. *Mathematical and Computer Modelling, 12*(12), 1613-1625. https://doi.org/10.1016/0895-7177(89)90337-3

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika, 64*(2), 191-199. https://doi.org/10.2307/2335684

Prescott, R. J. (2018). Editorial: Avoid being tripped up by statistics: Statistical guidance for a successful research paper. *Gait & Posture*, in-press. https://doi.org/10.1016/j.gaitpost.2018.06.172

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin, 86*(2), 420-428. https://doi.org/10.1037/0033-2909.86.2.420

Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics, 16*(2), 117-186. https://doi.org/10.1214/aoms/1177731118

Wang, S. K., & Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics, 43(1)*, 193-199. https://doi.org/10.2307/2531959

Appendix 1

The R code for the simulation study in Section 3.4 is provided below.

```
m = 3
n = 1000
alpha = 0.05
power = 0.95
beta = 1 - power
mu = 100
sigma.true = 0.06
sigma0 = 0.05
sigma1 = 0.06
n.sim = 10000
N = D = rep( NA, n.sim )
c = ( sigma1 / sigma0 ) ^ 2
a = ( m - 1 ) / 2
b = 1 / ( 2 * c )
bound.lower = beta / ( 1 - alpha )
bound.upper = ( 1 - beta ) / alpha
for ( k in 1:n.sim ) {
   full = matrix( rnorm( n * m, mu, sigma.true ), nrow=n, ncol=m )
   samp.var = apply( full, 1, var )
   test.stat = ( m - 1 ) * samp.var / sigma0 ^ 2
   f0 = cumprod( dchisq( test.stat, m - 1 ) )
   f1 = cumprod( dgamma( test.stat, a, b ) )
   sprt.stat = f1 / f0
   index = which( sprt.stat <= bound.lower | sprt.stat >= bound.upper )
   index = c( index, n )
   N[k] = temp = min( index, na.rm=TRUE )
   D[k] = ifelse( sprt.stat[temp] > 1, 1, 0 )
   }
mean(N)
mean(D)
```

Appendix 2

The R code for the simulation study in Section 4.4 is provided below.

```
m = 5
n = 1000
alpha = 0.05
power = 0.95
beta = 1 - power
sigma0.true = 1
sigma1.true = 1.2
tau0 = 1
tau1 = 1.2
n.sim = 10000
N = D = rep( NA, n.sim )
c = ( tau1 / tau0 ) ^ 2
bound.lower = beta / ( 1 - alpha )
bound.upper = ( 1 - beta ) / alpha
for ( k in 1:n.sim ) {
   y0 = matrix( rnorm( n * m, 0, sigma0.true ), nrow=n, ncol=m )
   y1 = matrix( rnorm( n * m, 0, sigma1.true ), nrow=n, ncol=m )
   samp0.var = apply( y0, 1, var )
   samp1.var = apply( y1, 1, var )
   test.stat = samp1.var / samp0.var * ( 1 / tau0 ) ^ 2
   f0 = cumprod( df( test.stat, m - 1, m - 1 ) )
   f1 = cumprod( df( test.stat / c, m - 1, m - 1 ) / c )
   sprt.stat = f1 / f0
   index = which( sprt.stat <= bound.lower | sprt.stat >= bound.upper )
   index = c( index, n )
   N[k] = temp = min( index, na.rm=TRUE )
   D[k] = ifelse( sprt.stat[temp] > 1, 1, 0 )
   }
mean(N)
mean(D)
```

Appendix 3

The R code for the simulation study in Section 5.4 is provided below.

```
m = 5
n = 1000
alpha = 0.05
power = 0.95
beta = 1 - power
sigma0 = 0.16
sigma1 = 0.2
theta0 = 0
theta1 = 0.1
theta.true = 0.1
n.sim = 10000
N = D = rep( NA, n.sim )
delta = sqrt(m) * ( theta1 - theta0 ) / sqrt( sigma1 ^ 2 + sigma0 ^ 2 )
bound.lower = beta / ( 1 - alpha )
bound.upper = ( 1 - beta ) / alpha
for ( k in 1:n.sim ) {
   full = matrix( rnorm( n * m, theta.true, sqrt( sigma1 ^ 2 + sigma0 ^ 2 ) ),
                  nrow=n, ncol=m )
   samp.var = apply( full, 1, var )
   samp.mean = apply( full, 1, mean )
   test.stat = ( samp.mean - theta0 ) / sqrt( samp.var / m )
   f0 = cumprod( dt( test.stat, m - 1 ) )
   f1 = cumprod( dt( test.stat, m - 1, delta ) )
   sprt.stat = f1 / f0
   index = which( sprt.stat <= bound.lower | sprt.stat >= bound.upper )
   index = c( index, n )
   N[k] = temp = min( index, na.rm=TRUE )
   D[k] = ifelse( sprt.stat[temp] > 1, 1, 0 )
   }
mean(N)
mean(D)
```

Appendix 4

To consider the case when the normality assumption is invalid, measurement error ϵ is generated from a Gumbel distribution with the PDF

$$f(\epsilon) = \frac{1}{\gamma_2} \exp\left\{-\left(e^{-\epsilon} + \frac{\epsilon - \gamma_1}{\gamma_2}\right)\right\}$$

for $-\infty < \gamma_1 < \infty$ and $\gamma_2 > 0$ and a mixed exponential (ME) distribution with the PDF

$$f(\epsilon) = w \frac{1}{\gamma_1} e^{-\epsilon/\gamma_1} I_{\epsilon \geq 0} + (1 - w) \frac{1}{\gamma_2} e^{\epsilon/\gamma_2} I_{\epsilon < 0}$$

for $\gamma_1 > 0$, $\gamma_2 > 0$, and $0 < w < 1$. The parameter values are chosen to match with the simulation scenarios in Section 5.4. For the ME distribution, the parameter values are gradually changed to test different skewness (while preserving the effect size in each simulation scenario).

Table 10. Probability of rejecting H_0 in SPRT designed for $\theta_0 = 0$, $\theta_1 = 0.1$, $\sigma_0 = 0.16$, $\sigma_1 = 0.20$, $\alpha = 0.05$ and $1 - \zeta = 0.95$ when errors are generated from normal and other distributions

Error Model	Skewness	θ	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$	$m = 8$	$m = 9$	$m = 10$
Normal	0.00	0.00	0.044	0.038	0.035	0.033	0.036	0.029	0.030	0.024	0.028
		0.10	0.962	0.964	0.966	0.968	0.968	0.970	0.973	0.973	0.974
Gumbel	1.14	0.00	0.024	0.023	0.023	0.024	0.026	0.020	0.021	0.020	0.019
		0.10	0.967	0.971	0.971	0.973	0.972	0.976	0.976	0.976	0.977
ME	0.00	0.00	0.040	0.036	0.031	0.036	0.033	0.027	0.029	0.027	0.025
		0.10	0.987	0.983	0.981	0.982	0.978	0.978	0.980	0.982	0.982
ME	1.85	0.00	0.021	0.020	0.020	0.022	0.020	0.018	0.018	0.014	0.017
		0.10	0.993	0.991	0.987	0.988	0.987	0.987	0.984	0.985	0.985
ME	5.66	0.00	0.007	0.008	0.006	0.006	0.007	0.006	0.008	0.007	0.006
		0.10	1.000	1.000	1.000	0.999	0.999	0.998	0.998	0.997	0.995

Table 11. Average sample size $E(N)$ in SPRT designed for $\theta_0 = 0$, $\theta_1 = 0.1$, $\sigma_0 = 0.16$, $\sigma_1 = 0.20$, $\alpha = 0.05$ and $1 - \zeta = 0.95$ when errors are generated from normal and other distributions

Error Model	Skewness	θ	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$	$m = 8$	$m = 9$	$m = 10$
Normal	0.00	0.00	24.4	15.8	11.8	9.5	7.9	6.9	6.1	5.5	5.0
		0.10	25.4	16.1	12.2	9.8	8.2	7.1	6.3	5.7	5.2
Gumbel	1.14	0.00	22.2	14.4	11.0	9.0	7.5	6.6	5.9	5.3	4.9
		0.10	24.5	15.7	11.9	9.5	8.0	7.0	6.1	5.5	5.1
ME	0.00	0.00	24.9	15.9	11.7	9.3	8.0	6.8	6.1	5.5	5.0
		0.10	21.0	14.1	10.8	8.8	7.5	6.6	5.8	5.4	4.9
ME	1.85	0.00	21.4	13.8	10.6	8.5	7.3	6.4	5.7	5.2	4.8
		0.10	17.4	12.4	9.7	8.1	6.9	6.2	5.5	5.1	4.6
ME	5.66	0.00	17.4	11.6	8.8	7.3	6.2	5.5	5.0	4.6	4.3
		0.10	8.7	6.2	5.1	4.4	3.9	3.6	3.4	3.2	3.0