Combining Standardized Mean Differences Using the Method of Maximum Likelihood

Ke-Hai Yuan
University of Notre Dame
and
Brad J. Bushman
Iowa State University

May 1, 2000; Revised, May 21, 2001
Second revision, January 28, 2002

Abstract

A maximum likelihood procedure for combining standardized mean differences based on a noncentral t-distribution is proposed. With a proper data augmentation technique, an EM-algorithm is developed. Information and likelihood ratio statistics are discussed in detail for reliable inference. Simulation results favor the proposed procedure over both the existing normal theory maximum likelihood procedure and the commonly used generalized least squares procedure.

Key words: Noncentral t-distribution, data augmentation, EM-algorithm, observed information, Monte-Carlo, meta-analysis, effect size, maximum likelihood, weighted least squares.

1. Introduction

As the number of scientific studies continues to grow at an exponential rate, it becomes increasingly important to integrate the results from these studies. It is somewhat ironic that the traditional review of scientific research has been conducted in an unscientific fashion. In the traditional narrative (qualitative) review, the reviewer uses "mental algebra" to combine the findings from a collection of studies, and describes the results verbally. Narrative reviews are particularly susceptible to the subjective judgments, preferences, and biases of a particular reviewer's perspective. Statisticians were the first to advocate a more scientific approach to reviewing the literature - an approach based on quantitative methods. These quantitative methods were labeled meta-analysis by Glass (1976, p. 3):

> Meta-analysis refers to the analysis of analyses the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating findings. It connotes a rigorous alternative to the casual, narrative discussions of research studies which typify our attempts to make sense of the rapidly expanding literature.

In the meta-analytic (quantitative) review, the reviewer uses statistical procedures to integrate the findings from a collection of studies, and describes the results using a numerical effect-size estimate (e.g., Wang & Bushman, 1999). An effect size provides a numerical measure of how strongly two variables are related (e.g., viewing violence and behaving aggressively, smoking tobacco and getting lung cancer). When the primary studies compare two groups, either through experimental (treatment) versus control comparisons or through planned contrasts, the effect size estimate is expressed as some form of standardized difference between group means. This article focuses on estimating the population standardized mean difference. Hedges (1980) proposed using the method of maximum likelihood for estimating the standardized mean difference. His estimate is based on the normal density functions for the experimental and control groups. Because of many nuisance parameters involved in this likelihood function (in addition to the mean difference of interest), the resulting estimator may not be consistent even when many studies are combined (Hedges & Olkin, 1985). Consequently, most meta-analyses use a weighted least squares estimator. In this article we propose an alternative maximum likelihood estimator based on the noncentral t-distribution. First, we briefly review the commonly used weighted least squares (WLS) method (Hedges, 1982), and the normal theory based maximum likelihood method (Hedges, 1980, 1982). Next, we describe an alternative method for estimating the standardized mean difference.

Suppose that the data for a meta-analysis arise from a series of k independent studies, each of which compares a treatment or experimental group (e) with a control group (c). Let y_{eij} and y_{cij} denote the jth observations in the experimental and control groups of the ith study, and let n_{ei} and n_{ci} be the experimental and control group sample sizes. Then it is typical to assume

$$y_{eij} \sim N(\mu_{ei}, \sigma_i^2),\ j = 1, \cdots n_{ei};\ \text{ and }\ y_{cij} \sim N(\mu_{ci}, \sigma_i^2),\ j = 1, \cdots, n_{ci}. \tag{1}$$

Let

$$\bar{y}_{ei} = \sum_{j=1}^{n_{ei}} y_{eij}/n_{ei}, \quad \bar{y}_{ci} = \sum_{j=1}^{n_{ci}} y_{cij}/n_{ci},$$

$$s_{ei}^2 = \sum_{j=1}^{n_{ei}} (y_{eij} - \bar{y}_{ei})^2/(n_{ei} - 1), \quad \text{and} \quad s_{ci}^2 = \sum_{j=1}^{n_{ci}} (y_{cij} - \bar{y}_{ci})^2/(n_{ci} - 1).$$

Glass (1976) proposed a simple approach of combining standardized mean differences $\delta_i = (\mu_{ei} - \mu_{ci})/\sigma_i$. Define the pooled estimate of variance as

$$s_i^2 = [(n_{ei} - 1)s_{ei}^2 + (n_{ci} - 1)s_{ci}^2]/(n_{ei} + n_{ci} - 2),$$

the standardized mean difference δ_i can be estimated by

$$d_i = \frac{\bar{y}_{ei} - \bar{y}_{ci}}{s_i}, \tag{2}$$

however, d_i is well known a biased estimator of δ_i. Hedges (1981) proposed an unbiased estimator of δ_i,

$$d_i^u = c(p_i)d_i, \tag{3}$$

where $c(p_i) = \Gamma(p_i/2)/\{\sqrt{p_i/2}\Gamma[(p_i - 1)/2]\}$ and $p_i = n_{ei} + n_{ci} - 2$. Hedges (1982) further proposed combining effect sizes using the method of WLS. When the combined sample size is large (i.e., p_i is large), Hedges (1982) suggested the following approximation to the distribution of d_i^u:

$$d_i^u \sim N(\delta_i, \sigma_i^2(\delta_i)), \tag{4}$$

where

$$\sigma_i^2(\delta_i) = \frac{n_{ei} + n_{ci}}{n_{ei} n_{ci}} + \frac{\delta_i^2}{2(n_{ei} + n_{ci})}.$$

Let

$$c_p = 1 - \frac{3}{4p - 1},$$

then

$$c(p) \approx c_p.$$

Substituting c_{p_i} for $c(p_i)$ in equation (3) yields the commonly used procedure given by Hedges (1982) for combining standardized mean differences

$$\hat{\delta}_{WLS} = \frac{\sum_{i=1}^{k} w_i d_i^u}{\sum_{i=1}^{k} w_i} \quad \text{with} \quad \mathrm{Var}(\hat{\delta}_{WLS}) = \frac{1}{\sum_{i=1}^{k} w_i}, \tag{5}$$

where $w_i = 1/\sigma_i^2(d_i^u)$. The estimator $\hat{\delta}_{WLS}$ is used to estimate the common effect size δ under the null hypothesis

$$H_0: \quad \delta_1 = \cdots = \delta_k = \delta. \tag{6}$$

It is easy to see that the solution $\hat{\delta}_{WLS}$ corresponds to minimizing the WLS function

$$Q(\delta) = \sum_{i=1}^{k} w_i (d_i^u - \delta)^2.$$

Hedges (1982) and Rosenthal and Rubin (1982) independently proposed the test statistic $Q(\hat{\delta}_{WLS})$ for testing the null hypothesis given in equation (6). Under the null hypothesis, $Q(\hat{\delta}_{WLS}) \sim \chi_{k-1}^2$.

According to classical statistical theory, the maximum likelihood approach is usually preferred over the WLS approach, unless the data are nonnormally distributed. The advantages of the maximum likelihood estimator (MLE) are that it is asymptotically most efficient and its distribution is easily obtained. Furthermore, the likelihood ratio statistic asymptotically follows a chi-square distribution. These properties are well known (e.g., Stuart & Ord, 1991) and are lacking for other type of estimators.

In order to obtain a MLE for δ, Hedges (1980, 1982) formulated the following statistical model for the effect size δ under the null hypothesis in equation (6):

$$y_{eij} = \delta\sigma_i + \gamma_i + \epsilon_{eij}, \quad j = 1, \cdots, n_{ei}, \tag{7a}$$

$$y_{cij} = \gamma_i + \epsilon_{cij}, \quad j = 1, \cdots, n_{ci}. \tag{7b}$$

Based on the sufficient statistics $(\bar{y}_{ei}, s^2_{ei}, \bar{y}_{ci}, s^2_{ci})$ for model (7), a MLE of δ can be obtained by solving

$$\sum_{i=1}^{k} \tilde{n}_i[\tilde{n}_i \frac{(\bar{y}_{ei} - \bar{y}_{ci})^2}{h_i} - 2]\delta + \sum_{i=1}^{k} \text{sign}[(\bar{y}_{ei} - \bar{y}_{ci})]\tilde{n}_i^2 \frac{(\bar{y}_{ei} - \bar{y}_{ci})^2}{h_i}(\delta^2 + g_i)^{1/2} = 0, \tag{8}$$

which is the equation (4.3) of Hedges (1980), where $\tilde{n}_i = n_{ci}n_{ei}/(n_{ci} + n_{ei})$,

$$h_i = p_i s_i^2 + \tilde{n}_i(\bar{y}_{ei} - \bar{y}_{ci})^2 \quad \text{and} \quad g_i = 4(n_{ci} + n_{ei})a_i/[\tilde{n}_i^2(\bar{y}_{ci} - \bar{y}_{ei})^2].$$

However, as discussed in Hedges and Olkin (1985), there are two major disadvantages with the MLE based on model (7). First, with a single effect size the MLE $d_i' = \sqrt{p_i + 2}d_i/\sqrt{p_i}$ is positively biased. Second, the MLE based on model (7) may not be consistent if n_{eij} and n_{cij} remain small even when $k \to \infty$. This is because the number of nuisance parameters (σ_i, γ_i) increases as k increases (Neyman & Scott, 1948). Due to these two disadvantages, one might wonder whether the MLE based on model (7) is the best estimator. As we shall see, simulation results show that the estimator obtained by solving (8) is positively biased even when k is fairly large.

The likelihood function formulated through model (7) contains nuisance parameters (γ_i, σ_i). If a likelihood function can be derived that does not contain nuisance parameters, then the Neyman-Scott type of problems associated with MLE can be resolved. The positive bias associated with d_i' can also be resolved if a proper model for δ based on model (1) can be derived. The objective of this paper is to develop an approach for a MLE of δ using a different model formulation.

If $m_i = [n_{ei}n_{ci}/(n_{ei} + n_{ci})]^{1/2}$, then under models (1) and (6) we have

$$x_i = m_i d_i \sim t_{p_i}(\lambda_i), \tag{9}$$

where $\lambda_i = m_i\delta$. It is obvious that besides δ there is no other unknown parameter in the density function of the noncentral t-distribution in equation (9). Thus, the MLE based on model (9) avoids the drawbacks associated with the MLE based on model (7). As we shall see, the MLE based on model (9) also possesses nice small sample properties and is a competitive alternative to $\hat{\delta}_{WLS}$. Of course, the density function of a noncentral t-distribution is much more complicated than is the density function of a normal distribution. However, by using proper data augmentation and applying an EM-algorithm, one can obtain the MLE based on model (9) without any numerical difficulty.

In section 2 we describe an EM-algorithm based on data augmentation. In section 3 we discuss inference issues regarding the population standardized mean difference. In section 4 we describe the results of a simulation study that compares estimates based on the three methods. In section 5 we present some empirical evidence regarding the convergence of the EM-algorithm and the stability of the three estimators when an extreme effect size is included. In section 6 we discuss the EM-algorithm and offer some conclusions.

2. Data Augmentation and EM-Algorithm

If $y \sim N(\lambda, 1)$ and $u \sim \chi_p$, then we can write $x \sim t_p(\lambda)$ as

$$x = y\sqrt{p}/u.$$

In a meta-analysis, x is observed, but y and u are not observed. We will call y and u latent variables. In order to develop an EM-algorithm for estimating λ, the observed data x are augmented with the latent variable u. Because the conditional distribution of x given u is

$$(x|u) \sim N(\lambda\sqrt{p}/u, p/u^2),$$

the augmented likelihood function is

$$l(\lambda|x, u) = f(x|u)f(u) = \frac{u^p}{2^{p/2-1}\Gamma(p/2)\sqrt{2\pi p}} \exp[-\frac{u^2}{2} - \frac{u^2}{2p}(x - \lambda\sqrt{p}/u)^2]. \tag{10}$$

With a single effect size, our aim is to derive an estimator $\hat{\delta}_{ML}$ through $\hat{\lambda}$ that satisfies the condition

$$l(\hat{\lambda}|x) = \max_{\lambda} l(\lambda|x). \tag{11}$$

However, the form of the likelihood function $l(\lambda|x)$ for a noncentral t-distribution is so complicated that a direct solution to equation (11) is formidable if not impossible. An EM-algorithm offers an alternative and easy to implement approach to solve for $\hat{\lambda}$. The EM-algorithm was first proposed by Dempster, Laird and Rubin (1977), and it has been used in a great number of statistical applications (e.g., McLachlan & Krishnan, 1997; Tanner, 1996).

Based on the augmented likelihood function (10), the EM-algorithm is an iterative algorithm consisting of two steps: the E-step and the M-step. Let $\lambda^{(j)}$ be the current estimate of λ and let $f(u|x, \lambda^{(j)})$ be the conditional density of u given x. The E-step is to calculate

$$Q(\lambda, \lambda^{(j)}) = \int_0^\infty \log[f(\lambda|x, u)]f(u|x, \lambda^{(j)})du.$$

The M-step is to maximize Q with respect to λ at $\lambda^{(j+1)}$, treating $\lambda^{(j)}$ as fixed. The process is repeated until $|\lambda^{(j+1)} - \lambda^{(j)})|$ is sufficiently small.

Before giving more details of the EM-algorithm, we want to point out that the MLE $\hat{\delta}_{ML}$ defined using equation (11) is different from $d_i' = \sqrt{p_i + 2}d_i/\sqrt{p_i}$ for a single standardized mean difference. This can be illustrated using an example.

Example 1. Let $\delta = (\mu_e - \mu_c)/\sigma$ be a single standardized mean difference. Let $n_e = n_c = 5$. Suppose we observe $d = 1.0$, where d is defined in equation (2). Then $x = (5/2)^{1/2}$, and the MLE based on model (7) is $d' = \sqrt{5/4} \approx 1.118$; $\hat{\delta}_{WLS} = .903$; and the MLE defined using equation (11) is $\hat{\delta}_{ML} = 1.028$. The likelihood function $l(\delta|x = \sqrt{5/2})$ of $x \sim t_1(\sqrt{5/2}\delta)$ is a concave function of δ as can be seen in Figure 1.

Insert Figure 1 about here

Suppose a meta-analysis includes k effect sizes. Following the notation used in equation (9), the observed data are $\mathbf{x} = (x_1, \cdots, x_k)'$. Each x_i follows a noncentral t-distribution with degrees of freedom p_i and noncentrality parameter $\lambda_i = m_i\delta$. Let u_i be a chi-distribution with degrees of freedom p_i, where the u_i, $i = 1, \cdots, k$ are independent. If $\mathbf{u} = (u_1, \cdots, u_k)'$, then our augmented data are $(\mathbf{x}, \mathbf{u})$. After dropping a constant term, the joint log likelihood function of $(\mathbf{x}, \mathbf{u})$ is

$$ll(\delta|\mathbf{x}, \mathbf{u}) = -\sum_{i=1}^{k} \frac{x_i^2 + p_i}{2p_i} u_i^2 + \sum_{i=1}^{k} \frac{x_i u_i m_i \delta}{\sqrt{p_i}} - \frac{\delta^2}{2} \sum_{i=1}^{k} m_i^2 + \sum_{i=1}^{k} p_i \ln(u_i). \tag{12}$$

If $u_i^{(j)} = E(u_i|x_i, \delta^{(j)})$, then the EM-algorithm for obtaining $\hat{\delta}_{ML} = \delta^{(\infty)}$ is given by

$$\delta^{(j+1)} = \sum_{i=1}^{k} \frac{x_i m_i}{\sqrt{p_i}} u_i^{(j)} / \sum_{i=1}^{k} m_i^2. \tag{13}$$

At each iteration one calculates $u_i^{(j)}$, which is the conditional expectation of u_i given $(x_i, \delta^{(j)})$. After simplification, the conditional density function of u_i given $(x_i, \delta^{(j)})$ can be written as

$$f_i(u_i|x_i, \delta^{(j)}) = c_i u^{p_i} \exp\{-\frac{(x_i^2 + p_i)}{2p_i}[u_i - \frac{x_i\sqrt{p_i}m_i\delta_i^{(j)}}{(x_i^2 + p_i)}]^2\}$$

with

$$c_i^{-1} = \int_0^\infty u^{p_i} \exp\{-\frac{(x_i^2+p_i)}{2p_i}[u_i - \frac{x_i\sqrt{p_i}m_i\delta_i^{(j)}}{(x_i^2+p_i)}]^2\}du_i. \quad (14)$$

Various numerical approaches are available for evaluating

$$u_i^{(j)} = \int_0^\infty u_i f_i(u_i|x_i, \delta^{(j)})du_i. \quad (15)$$

But with the appropriate statistical software, one does not need to actually evaluate these integrals in equations (14) and (15). Notice that the integrals in equations (14) and (15) are of the form

$$A_p(a,b) = \int_0^\infty u^p \exp[-a(u-b)^2]du. \quad (16)$$

When $b > 0$, using substitution $u - b = v$ we have

$$A_p(a,b) = A_{p1}(a,b) + A_{p2}(a,b),$$

where

$$A_{p1}(a,b) = \int_0^\infty (v+b)^p \exp(-av^2)dv, \quad A_{p2}(a,b) = \int_{-b}^0 (v+b)^p \exp(-av^2)]dv.$$

Further calculation leads to

$$A_{p1}(a,b) = \frac{1}{2}\sum_{q=0}^{p} \frac{\Gamma(p+1)}{\Gamma(q+1)\Gamma(p-q+1)} \frac{b^{(p-q)}}{a^{(q+1)/2}}\Gamma(\frac{q+1}{2}),$$

and

$$A_{p2}(a,b) = \frac{1}{2}\sum_{q=0}^{p}(-1)^q \frac{\Gamma(p+1)}{\Gamma(q+1)\Gamma(p-q+1)} \frac{b^{(p-q)}}{a^{(q+1)/2}} \int_0^{ab^2} v^{(q-1)/2}e^{-v}dv. \quad (17)$$

Almost every statistical software package includes the gamma function. The integral in equation (17) can be calculated using the incomplete gamma function that exists in both SAS and MATLAB software. For example, denote the incomplete gamma function

$$\Gamma_{\text{in}}(s,t) = \frac{1}{\Gamma(s)}\int_0^t x^{s-1}e^{-x}dx, \quad (18)$$

then

$$\int_0^{ab^2} v^{(q-1)/2}e^{-v}dv = \Gamma(\frac{q+1}{2})\Gamma_{\text{in}}(\frac{q+1}{2}, ab^2). \quad (19)$$

It follows from equations (17) and (19) that

$$A_{p2}(a,b) = \frac{1}{2}\sum_{q=0}^{p}(-1)^q \frac{\Gamma(p+1)}{\Gamma(q+1)\Gamma(p-q+1)} \frac{b^{(p-q)}}{a^{(q+1)/2}}\Gamma(\frac{q+1}{2})\Gamma_{\text{in}}(\frac{q+1}{2}, ab^2).$$

If $c = -b$ when $b < 0$, then we have

$$A_p(a,b) = \sum_{q=0}^{p}(-1)^{(p-q)}\frac{\Gamma(p+1)}{\Gamma(q+1)\Gamma(p-q+1)}\frac{c^{p-q}}{a^{(q+1)/2}}\int_{ac^2}^{\infty} v^{(q-1)/2}e^{-v}dv.$$

Using the incomplete gamma function (18) we have

$$A_p(a,b) = \sum_{q=0}^{p}(-1)^{(p-q)}\frac{\Gamma(p+1)}{\Gamma(q+1)\Gamma(p-q+1)}\frac{c^{p-q}}{a^{(q+1)/2}}\Gamma(\frac{q+1}{2})[1-\Gamma_{\text{in}}(\frac{q+1}{2}, ac^2)].$$

If $a_i = (x_i^2 + p_i)/(2p_i)$ and $b_{ij} = x_i\sqrt{p_i}m_i\delta^{(j)}/(x_i^2+p_i)$, then we can express

$$u_i^{(j)} = A_{p_i+1}(a_i, b_{ij})/A_{p_i}(a_i, b_{ij}). \quad (20)$$

Equations (13) and (20) give the EM-algorithm for obtaining $\hat{\delta}_{ML}$ without difficulty using any program that includes the gamma and incomplete gamma functions.

3. Information and Likelihood Ratio Statistic

One advantage of the MLE is that its asymptotic distribution can be characterized by the associated information $\mathcal{I}$. In the present context, $\mathcal{I}$ is given by the negative expectation of the second derivative of the log likelihood function (12). Specifically, the v in

$$(\hat{\delta}_{ML} - \delta_0) \xrightarrow{\mathcal{L}} N(0, v) \quad (21)$$

is given by $\mathcal{I}^{-1}$. In the EM-algorithm, the information $\mathcal{I}$ is not a default output. One has to specifically calculate it. There are several approaches to calculate an information at the convergence of an EM-algorithm (e.g., McLachlan & Krishnan, 1997). The direct evaluation of the second derivative of the likelihood function appears to be the best approach. Denote

$$g_i(u_i) = \exp\{-\frac{(x_i^2+p_i)}{2p_i}[u_i - \frac{x_i\sqrt{p_i}m_i\delta}{(x_i^2+p_i)}]^2\}, \text{ and } A_i(\delta) = \int_0^{\infty} u^{p_i}g_i(u_i)du_i.$$

Then, after omitting a constant, the log likelihood function of (9) can be written as

$$ll_i(\delta) = -\frac{p_i m_i^2\delta^2}{2(x_i^2+p_i)} + \ln[A_i(\delta)]. \quad (22)$$

For convenience we will use dot on top of a function to denote a derivative. It follows from equation (22) that

$$\ddot{ll}_i(\delta) = -\frac{p_i m_i^2}{(x_i^2+p_i)} + \frac{\ddot{A}_i(\delta)}{A_i(\delta)} - \frac{\dot{A}_i^2(\delta)}{A_i^2(\delta)}, \quad (23)$$

where

$$\dot{A}_i(\delta) = \frac{x_i m_i}{\sqrt{p_i}} \int_0^\infty u^{p_i+1} g_i(u_i) du_i - \frac{x_i^2 m_i^2 \delta}{x_i^2 + p_i} \int_0^\infty u^{p_i} g_i(u_i) du_i,$$

and

$$\begin{aligned} \ddot{A}_i(\delta) &= [\frac{x_i^4 m_i^4 \delta^2}{(x_i^2+p_i)^2} - \frac{x_i^2 m_i^2}{x_i^2+p_i}] \int_0^\infty u^{p_i} g_i(u_i) du_i \\ &- 2\frac{x_i^3 m_i^3 \delta}{\sqrt{p_i}(x_i^2+p_i)} \int_0^\infty u^{p_i+1} g_i(u_i) du_i + \frac{x_i^2 m_i^2}{p_i} \int_0^\infty u^{p_i+2} g_i(u_i) du_i. \end{aligned}$$

Let $a_i = (x_i^2 + p_i)/(2p_i)$ and $b_i = x_i\sqrt{p_i} m_i \delta/(x_i^2 + p_i)$. Then with $A_p(a, b)$ as defined in equation (16) we have

$$\dot{A}_i(\delta) = \frac{x_i m_i}{\sqrt{p_i}} A_{p_i+1}(a_i, b_i) - \frac{x_i^2 m_i^2 \delta}{x_i^2 + p_i} A_{p_i}(a_i, b_i),$$

and

$$\ddot{A}_i(\delta) = [\frac{x_i^4 m_i^4 \delta^2}{(x_i^2+p_i)^2} - \frac{x_i^2 m_i^2}{x_i^2+p_i}] A_{p_i}(a_i, b_i) - 2\frac{x_i^3 m_i^3 \delta}{\sqrt{p_i}(x_i^2+p_i)} A_{p_i+1}(a_i, b_i) + \frac{x_i^2 m_i^2}{p_i} A_{p_i+2}(a_i, b_i).$$

Evaluation of the information $\mathcal{I}_i = -\ddot{l}l(\hat{\delta}_{ML})$ is tedious but straightforward. Notice that

$$\mathcal{I} = -\sum_{i=1}^{k} \ddot{l}l_i(\hat{\delta}_{ML}) \quad (24)$$

is the observed information, which is generally more accurate for describing the distribution of $\hat{\delta}_{ML}$ in equation (21) (Efron & Hinkley, 1978). Before continuing our discussion on information we use an example to illustrate the roles of n_e and n_c in $\mathcal{I}$.

Example 2. This example compares the information when n_e is fixed and n_c increases versus the information when both n_e and n_c increase. For simplicity, we assume the observed standardized mean difference $d = 1.0$. With $n_e = 5$, the informations with a single effect size corresponding to values of n_c from 5 to 145 are plotted in Figure 2(a). The corresponding informations for $n_e = n_c$ from 5 to 75 are plotted in Figure 2(b). As can be seen from Figure 2(a), when n_e is fixed the information increases as n_c increases, but the gain of extra information is diminishing. On the other hand, the information increases linearly as $n_e = n_c$ increase (see Figure 2(b)).

Insert Figure 2 about here

In contrast to the MLE based on model (7), which may not be consistent as $k \to \infty$ (Hedges & Olkin, 1985), the $\hat{\delta}_{ML}$ based on model (9) approaches δ_0 as $k \to \infty$ even when n_{ei} and n_{ci} are small. This can be established under fairly standard conditions (see Yuan & Jennrich, 1998). However, one does not need to assume that k approaches infinity in order for the result in equation (21) to hold. This is because $\mathcal{I}_i$ can be quite large when n_e and n_c are large enough, as illustrated in Example 2. Actually, one can write equation (21) as

$$\sqrt{\mathcal{I}}(\hat{\delta}_{ML} - \delta_0) \xrightarrow{\mathcal{L}} N(0,1). \tag{21a}$$

The $\mathcal{I}$ in equation (21a) plays the role of sample size. As long as $\mathcal{I}$ is large enough, the result in equation (21) will yield a reliable inference. For a single effect size, this is illustrated in the following example. From now on, we will denote the MLE defined in (8) by $\hat{\delta}_{ML}^{(old)}$ and the one defined in (11) by $\hat{\delta}_{ML}^{(new)}$.

Example 3. This example illustrates the distribution of $\hat{\delta}_{ML}^{(new)}$ as the information increases. We choose $\delta = 1.0$, $n_e = n_c = 10, 20, 40, 80$. For a single effect size, the estimates $\hat{\delta}_{ML}^{(new)}$ based on 500 simulation replications are plotted in Figure 3 using quantile-quantile (QQ) plots. Notice that, with a single study, both $\hat{\delta}_{ML}^{(old)}$ and $\hat{\delta}_{WLS}$ are multiples of the two-sample t-statistic, so they approach a normal distribution quickly. We also have the QQ plots of the corresponding $\hat{\delta}_{WLS}$ for comparison purposes, and the QQ plots of $\hat{\delta}_{ML}^{(old)}$ are identical to those of $\hat{\delta}_{WLS}$. As reflected by the QQ plots, the distributions of $\hat{\delta}_{ML}^{(new)}$ and $\hat{\delta}_{WLS}$ are quite similar, both approach normality as sample size increases. However, there are important differences between the three estimators that we will describe in section 4.

Insert Figure 3 about here

Let $\boldsymbol{\beta} = (\delta_1, \cdots, \delta_k)'$, where each δ_i is a free parameter. Then $\boldsymbol{\beta}$ represents the most general alternative hypothesis against the H_0 in equation (6). It is easy to see that $\boldsymbol{\beta}(\delta) = (\delta, \cdots, \delta)'$ corresponds to the null hypothesis, which is nested within $\boldsymbol{\beta}$. Based on this nesting, the likelihood ratio test statistic T_{LR} for H_0 is

$$T_{LR} = 2[ll(\hat{\boldsymbol{\beta}}_{ML}^{(new)}) - ll(\hat{\delta}_{ML}^{(new)})], \tag{25}$$

where $\hat{\boldsymbol{\beta}}_{ML}^{(new)} = (\hat{\delta}_{1ML}^{(new)}, \cdots, \hat{\delta}_{kML}^{(new)})'$ and $\hat{\delta}_{iML}^{(new)}$ is the MLE of δ_i based on the ith study. It is well known that a likelihood ratio statistic approximately follows a chi-square distribution when the sample size is large enough. Because the T_{LR} in (25) involves each $\hat{\delta}_{iML}^{(new)}$ that only depends on data from the ith experiment, we need to have all the $\mathcal{I}_i$ large in order for the T_{LR} to be well approximated by χ^2_{k-1}. Similarly, the $\mathcal{I}_i$ needs to be large in order for the T_{LR} to be well approximated by a noncentral chi-square when (6) does not hold.

4. Simulation Study

When sample size is large, the maximum likelihood method developed in section 2 has the advantage of yielding the most efficient parameter estimates for large sample sizes. One may wonder whether there is any advantage for the $\hat{\delta}_{ML}^{(new)}$ with small to medium sample sizes. We will compare the finite sample properties of the $\hat{\delta}_{ML}^{(new)}$ with those of $\hat{\delta}_{ML}^{(old)}$ and $\hat{\delta}_{WLS}$. We will focus on bias and efficiency for the three types of estimators.

In the simulation design we choose: δ_0 =.1, .2, $\cdots$, .9. Virtually all meta-analyses have found effect sizes in this range (Hedges, 1982). Based on empirical research, Cohen (1977) proposed $\delta_0 = 0.20$, 0.50, and 0.80 as small, medium, and large effects, respectively. So our effect size selection covers small to large effect sizes. The numbers of independent studies (k) in the meta-analysis were $k = 24$, 40, and 67. We obtained these values of k by counting the number of studies included in 390 meta-analytic studies published in *Psychological Bulletin* through the year 2000. The value $k = 24$ is the 25th percentile for these 390 meta-analyses, the value $k = 40$ is the 50th percentile, and the value $k = 67$ is the 75th percentile. The average sample sizes per group were 5, 10, 20, 50, and 100. Because not all sample sizes in the experimental and control groups are equal in typical social science research, we let n_{ei} and n_{ci} be generated using $n = [N \times U] + 2$, where U is a random variable that follows the uniform distribution on $(0, 2)$, and $[N \times U]$ is the integer part of $N \times U$. On average, $n_{ei} = n_{ci} = N + 2$ but the experimental and control group sample sizes generally are not equal for a specific experiment. For each combination of δ_0, k, n_{ei} and n_{ci}, 500 replications were performed. Let δ_r denote an effect size estimate in the rth replication, the bias and variance were calculated respectively as

$$\text{bias} = \bar{\delta} - \delta_0, \quad \text{and} \quad \text{var} = \sum_{r=1}^{500} (\delta_r - \bar{\delta})^2/500,$$

where $\bar{\delta} = \sum_{r=1}^{500} \delta_r/500$. The results corresponding to average sample sizes 5, 10, 20, 50, 100 are reported in Tables 1 to 5 respectively. From these tables, and more simulation conditions that not presented here, we notice that the $\hat{\delta}_{WLS}$ is essentially always negatively (or systematically) biased whereas the $\hat{\delta}_{ML}^{(old)}$ is essentially always positively (or systematically) biased. The amount of bias appears smaller with larger n_e and n_c but does not decrease when k increases. In contrast, the bias in $\hat{\delta}_{ML}^{(new)}$ is much smaller. Actually, unlike the biases in either $\hat{\delta}_{WLS}$ or $\hat{\delta}_{ML}^{(old)}$, the biases in $\hat{\delta}_{ML}^{(new)}$ alternate between positive and negative. Thus, the bias in $\hat{\delta}_{ML}^{(new)}$ may not be systematic. Equation (21a) implies that there is no systematic bias in $\hat{\delta}_{ML}^{(new)}$ when the information $\mathcal{I}$ is large.

Insert Tables 1 to 5 about here

Empirical standard deviations (SD) of the three estimators for each condition are also reported in Tables 1 to 5. Based on the fact that the SDs for $\hat{\delta}_{WLS}$ are less than those for either $\hat{\delta}_{ML}^{(new)}$ or $\hat{\delta}_{ML}^{(old)}$, one may be tempted to conclude that $\hat{\delta}_{WLS}$ is the most efficient one among the three. In the present context, however, this is a false conclusion. For a parameter estimator $\hat{\theta}$, another estimator $\tilde{\theta} = a\hat{\theta}$ always has a smaller standard deviation when $a < 1$. This does not imply that $\tilde{\theta}$ is a more efficient parameter estimator than is $\hat{\theta}$. Comparing SDs is not fair when one estimator is negatively biased. Instead of comparing SDs, we will compare coefficients of variation (CV) as a criterion for efficiency. The CV is given by

$$\text{CV} = \frac{\text{SD}}{\text{mean}}.$$

The CV for $\hat{\delta}_{WLS}$, $\hat{\delta}_{ML}^{(old)}$ and $\hat{\delta}_{ML}^{(new)}$ are reported in Tables 1 to 5. The corresponding CVs suggest that the three estimators are similar in terms of efficiency. Notice in Tables 1 to 5 that the standard deviations decrease when either k or the averaged sample size increases. The standard deviations increase as δ_0 increases, whereas the coefficients of variation decrease as δ_0 increases. This can be understood in the following sense: Measuring a larger effect size is easier than measuring a smaller effect size to a certain amount of relative accuracy.

5. Examples

It is well-known that the EM-algorithm can have a very slow rate of convergence. Accordingly, various acceleration methods have been developed (e.g., Jamshidian & Jennrich, 1993; Louis, 1982; Meilijson, 1989). For some empirical examples presented in Chapter 4 of

McLachlan and Krishnan (1997) the EM-algorithm may take more than 100 iterations to finally converge while an accelerated procedure takes far fewer (e.g., 11) iterations. In order to understand the convergence of the particular EM-algorithm in (13), we also apply it to several empirical examples. In addition to study the convergence of (13), we also compare the sensitivity of $\hat{\delta}_{WLS}$, $\hat{\delta}_{ML}^{(old)}$ and $\hat{\delta}_{ML}^{(new)}$ when an extreme effect size exists in the data to be combined.

Our data sets are from Tables 6 to 9 of Hedges and Olkin (1985, pp. 23–25), these are respectively the effect of open education on attitude toward school, on student independence and self-reliance, on student self-concept, and on student creativity, with $k = 11$, 7, 18 and 10 studies collected by Hedges, Giaconia, and Gage (1981). For obtaining $\hat{\delta}_{ML}^{(old)}$ by solving equation (8), we use the Newton-Raphson algorithm. Using $\hat{\delta}_{WLS}$ as the initial value in both the Newton-Raphson and the EM-algorithms, the number of iterations for convergence are presented in Table 6. Our convergence criterion was chosen as $|(\delta^{(j+1)} - \delta^{(j)})/\delta^{(j)}| < 10^{-8}$, where $\delta^{(j)}$ is the value of δ after the jth iteration.

Insert Table 6 about here

The Newton Raphson algorithm for $\hat{\delta}_{ML}^{(old)}$ converges with 3, 2, 2, 3 iterations respectively for the four data sets whereas the EM-algorithm for $\hat{\delta}_{ML}^{(new)}$ converges in 5, 5, 4 and 6 iterations respectively. Contrasting to the accelerated EM-algorithms reported in McLachlan and Krishnan (1997), the EM-algorithm in the context for combining standardized mean differences in (13) probably will not need any accelerations. Actually, the results in Table 6 were calculated on a PC with a Pentium II 400 processor, which was considered a rather slow computer, and each row was finished instantly. So the slower convergence of (13) is not a serious issue for typical data sets in practice. The convergence of (13) in a few iterations is probably due to the fact that $\hat{\delta}_{WLS}$ is a very good starting value and there is only one unknown parameter δ involved.

In addition to the number of iterations in Table 6, we have also reported the $\hat{\delta}_{WLS}$, $\hat{\delta}_{ML}^{(old)}$ and $\hat{\delta}_{ML}^{(new)}$ together with their standard errors. The standard error of $\hat{\delta}_{WLS}$ is given by $(1/\sum_{i=1}^{k} w_i)^{1/2}$, the standard error of $\hat{\delta}_{ML}^{(old)}$ is calculated through the square root of the reciprocal of the information corresponding to δ in model (7) as given in equation (4.7) of

Hedges (1980), the standard error of $\hat{\delta}_{ML}^{(new)}$ is calculated according to the one developed in section 3.

In the rest of this section we will illustrate the effect of an extreme effect size on the three estimators $\hat{\delta}_{WLS}$, $\hat{\delta}_{ML}^{(old)}$ and $\hat{\delta}_{ML}^{(new)}$. For such a purpose, we take the data set in Table 6 of Hedges and Olkin (1985, p. 24) and add an extreme study effect to it. So the total number of studies combined is $k = 12$ instead of the original $k = 11$. First, we fix the sample sizes at $n_{e12} = n_{c12} = 2$ while changing the standardized mean difference from $d_{12} = 1$ to $d_{12} = 5$. All of these values exceed Cohen's (1977) conventional value for a large effect (i.e., $d = 0.80$). The corresponding changes in the three estimates are shown in the first five rows of Table 7. All three estimators are quite robust to extreme outlier, each only increases by about .005 when adding $d_{12} = 5$. A similar pattern of change in the reverse direction can be observed when adding $d_{12} = -1$ to $d_{12} = -5$, as shown in the next five rows of Table 7.

Insert Table 7 about here

The lower portion of Table 7 contains the result when fixing $d_{12} = 3$ and $d_{12} = -3$ while changing sample sizes $n_{e12} = n_{c12} = 5$ to 80. Different from the top portion, each estimator in the lower portion of Table 7 is strongly influenced by the sample sizes associated with the extreme effect size. For example, when $n_{e12} = n_{c12} = 80$ and $d_{12} = -3$ the resulting estimator is less than half of the estimator corresponding to the original data set. When comparing the three estimators, the two MLE's have similar reactions to the changing sample size while the reaction of $\hat{\delta}_{WLS}$ is, to some degree, less sensitive.

When a study has small sample sizes n_e and n_c, it is likely that the effect size may not be reliable. Of course, outliers are more common in small samples than in large samples (e.g., think of getting 7 heads in 10 coin tosses versus 700 heads in 1,000 coin tosses). When n_e and n_c are relatively large, we may have to trust the corresponding effect size, although it is in the extreme range. The three methods of combining mean differences, especially the two MLEs, automatically adjust for outliers. So, in practice, we may not need to worry about extreme effect sizes so much if the effect sizes are combined using any of the three methods described in this article. Other more robust alternatives to the $\hat{\delta}_{MLE}$ are discussed in detail by Zhang and Schoeps (1997).

Table 7 also contains the numbers of iterations of the Newton-Raphson algorithm and the EM-algorithm, the later one converges within 10 iterations for all the cases.

6. Discussion and Conclusion

Since it was introduced by Dempster et al. (1977), the EM-algorithm has been applied to solving many practical problems that the traditional Newton type method cannot easily solve. One of its most important applications is to estimate the mean vector and covariance matrix of a multivariate normal distribution with missing data (Dempster et al., 1977; Little & Rubin, 1987). Rubin (1983) implemented the EM-algorithm for obtaining the means and covariances of a multivariate t-distribution with complete data. Little (1988) extended it to the case of missing data with a multivariate t-distribution of known degrees of freedom. Recently, Liu and Rubin (1995) further extended the EM-algorithm to the case of unknown degrees of freedom for a multivariate t-distribution. In this paper, we extended the EM-algorithm to the one dimensional noncentral t-distribution for estimating the δ in $t_{p_i}(m_i\delta)$. Although it has been fully realized that a rescaled version of the sample standardized mean difference follows a noncentral t-distribution, to our knowledge, the estimation of the common δ by maximizing the density function of a noncentral t-distribution has never been developed. Because the Newton-type method for directly maximizing the likelihood function is unusually complicated, we developed an EM-algorithm method instead. With a program language, the implementation of this algorithm is straightforward. As with any other applications of data augmentations, the value of the EM-algorithm in section 2 lies in its ability to solve complicated problems. This is the beauty of the tool of data augmentation.

The advantage of the new MLE is that it is asymptotically efficient. The variance of $\hat{\delta}_{ML}^{(new)}$ is given by the inverse of the associated information, which we discussed in detail in the context of combining standardized mean differences. The likelihood ratio statistic for testing the homogeneity of effect sizes is also straightforward. Compared to the old MLE based on model (7), the formulation of the new MLE does not involve any nuisance parameters and thus $\hat{\delta}_{ML}^{(new)}$ is still consistent even when the experimental and control group sample sizes are small. The approximation in (4) by Hedges (1982) is based on the assumption that the experimental and control group sample sizes are large in each study. When the experimental and control group sample sizes are small in each study, or even in some of the studies, the approximation in (4) is not very accurate and consequently the value of formula (5) is compromised. The only drawback with the new MLE is that it needs more computational time than is needed for $\hat{\delta}_{ML}^{(old)}$ or $\hat{\delta}_{WLS}$. However, this drawback is not really an issue in this age of modern computers. For example, for any of the 390 meta-analyses published in *Psychological Bulletin*, it will be an instant to obtain the $\hat{\delta}_{ML}^{(new)}$ on a desk top computer.

Notice that the approximation c_p of $c(p)$ was used in obtaining the $\hat{\delta}_{WLS}$. One may wonder if the negative bias in $\hat{\delta}_{WLS}$ is caused by this approximation. Actually, $c_p > c(p)$ for every p. Thus, if $c(p)$ is used instead of c_p in calculating $\hat{\delta}_{WLS}$, the bias will be even worse. Of course, the efficiency of the $\hat{\delta}_{WLS}$ will be the same whether c_p or $c(p)$ is used.

In summary, the amount of bias in $\hat{\delta}_{WLS}$ or $\hat{\delta}_{ML}^{(old)}$ increases as δ increases, it decreases when the average sample sizes n_e and n_c increase. The bias is insensitive to the number of combined studies k. The new MLE we propose is less biased than the old MLE and the WLS estimator, especially when the experimental and control group sample sizes are small. Moreover, the minimal bias that does exist in the new MLE is not systematic. Even though the bias in the other two estimators was small, it was systematic. Of course, unbiased estimators are more desirable than biased estimators. Thus, the new MLE should be a valuable tool to the practicing meta-analyst.

References

Cohen, J. J. (1977). *Statistical power analysis for the behavioral sciences (2nd ed.).* Hillsdale, NJ: Erlbaum.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *Journal of the Royal Statistical Society, B39*, 1–38.

Efron, B. & Hinkley, D. V. (1978). The observed versus the expected information. *Biometrika, 65*, 457–487.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3–8.

Hedges, L. V. (1980). *Combining the results of experiments using different scales of measurement.* Doctorial dissertation, Stanford, CA: Stanford University.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107–128.

Hedges, L. V. (1982). Estimating effect size from a series of independent experiments. *Psychological Bulletin, 92*, 490–499.

Hedges, L. V., Giaconia, R. M. & Gage, N. L. (1981). *The empirical evidence on the effectiveness of open education.* Stanford, CA: Stanford University School of Education.

Hedges, L. V. & Olkin, I. (1985). *Statistical methods for meta-analysis.* San Diego, CA: Academic Press.

Jamshidian, M. & Jennrich, R. I. (1993). Conjugate gradient acceleration of the EM algorithm. *Journal of the American Statistical Association, 88*, 221–228.

Little, R. J. A. (1988). Robust estimation of the mean and covariance matrix from data with missing values. *Applied Statistics*, *37*, 23–38.

Little, R. J. A. & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Liu, C. H. & Rubin, D. B. (1995). ML estimation of the multivariate t distribution with unknown degrees of freedom. *Statistica Sinica, 5*, 19–39.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, B44*, 226–233.

McLachlan, G. J. & Krishnan, T. (1997). *The EM algorithm and extensions.* New York: Wiley.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society, B51*, 127–138.

Neyman, J. & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrika, 16*, 1–32.

Rosenthal, R. & Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin, 92*, 500–504.

Rubin, D. B. (1983). Iteratively reweighted least squares. In N. L. Johnson & S. Kotz (eds.) *Encyclopedia of statistical sciences, Volume 4* (pp. 272–275). New York: Wiley.

Stuart, A. & Ord, J. K. (1991). *Kendall's advanced theory of statistics, Vol. 2 (5th ed.).* New York: Oxford University Press.

Tanner, M. A. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions (3rd ed.).* New York: Springer-Verlag.

Wang, M. C. & Bushman, B. J. (1999). Integrating results through meta-analytic review using SAS software. Cary, NC: SAS Institute Inc.

Yuan, K.-H. & Jennrich, R. I. (1998). Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis, 65*, 245–260.

Zhang, Z. & Schoeps, N. (1997). On robust estimation of effect size under semiparametric models. *Psychometrika, 62*, 201–214.

Table 1

Bias, Standard Deviation, and Coefficient of Variation of $\hat{\delta}_{WLS}$, $\hat{\delta}_{ML}^{(old)}$ and $\hat{\delta}_{ML}^{(new)}$

(Average sample sizes $n_e = 5$, $n_c = 5$)

		$\hat{\delta}_{WLS}$			$\hat{\delta}_{ML}^{(old)}$			$\hat{\delta}_{ML}^{(new)}$		
k	δ	bias×10^3	SD	CV	bias×10^3	SD	CV	bias×10^3	SD	CV
	.10	-2.712	0.122	1.250	18.742	0.150	1.259	8.847	0.136	1.252
	.20	-22.073	0.127	0.712	25.554	0.160	0.709	3.426	0.144	0.710
	.30	-32.652	0.136	0.507	29.808	0.166	0.502	0.799	0.152	0.504
	.40	-48.001	0.127	0.361	36.736	0.156	0.357	-2.839	0.142	0.359
24	.50	-41.341	0.132	0.288	54.550	0.160	0.288	10.211	0.147	0.288
	.60	-64.384	0.134	0.250	62.875	0.167	0.252	4.174	0.151	0.251
	.70	-75.046	0.131	0.210	57.995	0.160	0.211	-3.819	0.146	0.210
	.80	-80.952	0.145	0.201	90.350	0.179	0.202	9.959	0.163	0.201
	.90	-104.206	0.149	0.187	94.318	0.187	0.188	1.340	0.169	0.187
	.10	-11.379	0.103	1.167	9.495	0.127	1.157	-0.238	0.116	1.160
	.20	-29.649	0.101	0.593	11.034	0.125	0.594	-7.731	0.114	0.593
	.30	-27.837	0.103	0.377	32.705	0.125	0.377	4.620	0.115	0.377
	.40	-34.102	0.096	0.261	47.180	0.117	0.262	9.592	0.107	0.261
40	.50	-52.478	0.105	0.236	50.543	0.130	0.236	2.427	0.118	0.235
	.60	-63.726	0.106	0.198	71.526	0.134	0.199	8.543	0.121	0.199
	.70	-73.503	0.099	0.159	66.287	0.122	0.159	1.234	0.111	0.158
	.80	-95.370	0.108	0.153	75.323	0.136	0.155	-4.488	0.123	0.154
	.90	-108.610	0.105	0.133	88.462	0.132	0.133	-3.702	0.119	0.133
	.10	-12.370	0.078	0.888	9.238	0.098	0.893	-0.854	0.088	0.891
	.20	-20.635	0.078	0.437	19.217	0.096	0.439	0.706	0.088	0.439
	.30	-28.771	0.079	0.289	31.213	0.096	0.290	3.390	0.088	0.290
	.40	-46.587	0.079	0.223	38.438	0.099	0.225	-0.900	0.089	0.224
67	.50	-55.304	0.077	0.173	50.483	0.095	0.173	1.085	0.086	0.173
	.60	-67.313	0.075	0.141	59.897	0.093	0.140	0.812	0.084	0.140
	.70	-75.030	0.082	0.132	71.109	0.102	0.133	2.974	0.093	0.132
	.80	-87.295	0.077	0.109	77.341	0.096	0.109	0.501	0.087	0.109
	.90	-105.175	0.083	0.104	96.724	0.104	0.105	1.951	0.094	0.104

Table 2

Bias, Standard Deviation, and Coefficient of Variation of $\hat{\delta}_{WLS}$, $\hat{\delta}^{(old)}_{ML}$ and $\hat{\delta}^{(new)}_{ML}$

(Average sample sizes $n_e = 10$, $n_c = 10$)

		$\hat{\delta}_{WLS}$			$\hat{\delta}^{(old)}_{ML}$			$\hat{\delta}^{(new)}_{ML}$		
k	δ	bias$\times 10^3$	SD	CV	bias$\times 10^3$	SD	CV	bias$\times 10^3$	SD	CV
	.10	-3.967	0.096	1.000	5.267	0.105	0.999	1.116	0.101	0.999
	.20	-15.236	0.102	0.549	5.599	0.113	0.552	-3.871	0.108	0.550
	.30	-15.304	0.100	0.351	11.093	0.109	0.351	-0.759	0.105	0.351
	.40	-16.460	0.103	0.268	20.793	0.113	0.268	4.016	0.108	0.268
24	.50	-22.400	0.094	0.198	18.883	0.103	0.198	0.372	0.099	0.197
	.60	-37.867	0.097	0.172	18.387	0.106	0.172	-6.866	0.102	0.172
	.70	-34.628	0.089	0.135	23.041	0.098	0.135	-2.913	0.094	0.135
	.80	-31.310	0.099	0.129	42.213	0.109	0.129	8.940	0.104	0.129
	.90	-49.604	0.099	0.116	40.844	0.110	0.117	-0.325	0.105	0.116
	.10	-7.092	0.072	0.779	1.999	0.079	0.779	-2.087	0.076	0.779
	.20	-13.200	0.076	0.406	5.258	0.084	0.407	-3.079	0.080	0.407
	.30	-9.215	0.077	0.266	17.345	0.084	0.265	5.417	0.081	0.266
	.40	-14.296	0.071	0.184	21.318	0.077	0.184	5.367	0.075	0.184
40	.50	-29.243	0.078	0.165	15.202	0.085	0.165	-4.818	0.082	0.165
	.60	-34.039	0.077	0.136	26.938	0.085	0.136	-0.717	0.081	0.136
	.70	-30.736	0.072	0.107	30.750	0.078	0.107	2.947	0.075	0.107
	.80	-40.508	0.075	0.099	34.885	0.083	0.099	0.633	0.079	0.099
	.90	-46.634	0.077	0.091	39.959	0.085	0.091	0.531	0.082	0.091
	.10	-6.856	0.057	0.610	2.564	0.063	0.610	-1.686	0.060	0.610
	.20	-11.073	0.061	0.321	6.404	0.066	0.320	-1.445	0.064	0.320
	.30	-15.905	0.059	0.208	10.017	0.064	0.207	-1.608	0.062	0.207
	.40	-24.176	0.059	0.158	14.260	0.066	0.159	-3.104	0.063	0.158
67	.50	-21.694	0.059	0.124	25.129	0.065	0.124	3.975	0.063	0.124
	.60	-27.182	0.057	0.100	28.677	0.063	0.100	3.388	0.060	0.100
	.70	-39.908	0.057	0.086	24.719	0.063	0.086	-4.577	0.060	0.086
	.80	-39.372	0.058	0.077	33.426	0.064	0.077	0.459	0.062	0.077
	.90	-51.968	0.064	0.075	35.179	0.070	0.075	-4.501	0.067	0.075

Table 3

Bias, Standard Deviation, and Coefficient of Variation of $\hat{\delta}_{WLS}$, $\hat{\delta}_{ML}^{(old)}$ and $\hat{\delta}_{ML}^{(new)}$

(Average sample sizes $n_e = 20$, $n_c = 20$)

		$\hat{\delta}_{WLS}$			$\hat{\delta}_{ML}^{(old)}$			$\hat{\delta}_{ML}^{(new)}$		
k	δ	bias$\times 10^3$	SD	CV	bias$\times 10^3$	SD	CV	bias$\times 10^3$	SD	CV
	.10	0.112	0.070	0.695	4.442	0.073	0.696	2.540	0.071	0.695
	.20	-5.700	0.069	0.356	4.473	0.073	0.356	-0.025	0.071	0.356
	.30	-8.072	0.065	0.223	4.528	0.068	0.223	-1.025	0.067	0.223
	.40	-15.086	0.076	0.198	2.250	0.080	0.198	-5.403	0.078	0.198
24	.50	-9.398	0.073	0.149	10.214	0.076	0.149	1.590	0.075	0.149
	.60	-14.866	0.070	0.119	11.531	0.073	0.119	-0.157	0.071	0.119
	.70	-14.799	0.065	0.095	11.865	0.068	0.095	0.080	0.067	0.095
	.80	-23.312	0.070	0.090	10.810	0.073	0.090	-4.327	0.071	0.090
	.90	-18.519	0.075	0.085	24.910	0.079	0.086	5.506	0.077	0.085
	.10	-7.254	0.054	0.578	-2.916	0.056	0.577	-4.824	0.055	0.578
	.20	-4.625	0.056	0.286	4.259	0.059	0.286	0.332	0.057	0.286
	.30	-1.323	0.055	0.184	11.187	0.057	0.184	5.688	0.056	0.184
	.40	-9.392	0.056	0.143	6.845	0.058	0.143	-0.307	0.057	0.143
40	.50	-10.892	0.057	0.117	10.189	0.060	0.117	0.899	0.059	0.117
	.60	-19.778	0.057	0.098	8.989	0.060	0.098	-3.795	0.058	0.098
	.70	-20.831	0.054	0.079	8.002	0.056	0.079	-4.767	0.055	0.079
	.80	-19.228	0.054	0.069	16.855	0.057	0.069	0.799	0.056	0.069
	.90	-23.872	0.058	0.066	17.290	0.061	0.066	-1.073	0.059	0.066
	.10	-1.463	0.042	0.429	3.195	0.044	0.429	1.139	0.043	0.429
	.20	-5.874	0.042	0.214	2.298	0.043	0.214	-1.300	0.043	0.214
	.30	-8.341	0.040	0.137	3.778	0.042	0.137	-1.551	0.041	0.137
	.40	-10.434	0.043	0.109	8.004	0.045	0.109	-0.146	0.044	0.109
67	.50	-9.940	0.043	0.089	12.388	0.045	0.089	2.516	0.045	0.089
	.60	-15.887	0.043	0.074	10.211	0.045	0.074	-1.354	0.044	0.074
	.70	-17.592	0.043	0.063	13.172	0.045	0.063	-0.479	0.044	0.063
	.80	-20.845	0.042	0.053	13.045	0.043	0.053	-2.013	0.043	0.053
	.90	-22.257	0.043	0.049	19.374	0.045	0.049	0.811	0.044	0.049

Table 4

Bias, Standard Deviation, and Coefficient of Variation of $\hat{\delta}_{WLS}$, $\hat{\delta}_{ML}^{(old)}$ and $\hat{\delta}_{ML}^{(new)}$

(Average sample sizes $n_e = 50$, $n_c = 50$)

		$\hat{\delta}_{WLS}$			$\hat{\delta}_{ML}^{(old)}$			$\hat{\delta}_{ML}^{(new)}$		
k	δ	bias×10^3	SD	CV	bias×10^3	SD	CV	bias×10^3	SD	CV
	.10	-2.520	0.043	0.444	-0.886	0.044	0.444	-1.595	0.044	0.444
	.20	-2.268	0.046	0.235	1.729	0.047	0.235	-0.011	0.047	0.235
	.30	-1.839	0.043	0.144	3.075	0.044	0.145	0.941	0.043	0.144
	.40	-1.505	0.051	0.127	5.179	0.051	0.127	2.274	0.051	0.127
24	.50	-5.724	0.042	0.084	1.810	0.042	0.084	-1.463	0.042	0.084
	.60	-7.084	0.043	0.073	3.032	0.044	0.073	-1.390	0.044	0.073
	.70	-6.696	0.042	0.061	3.482	0.043	0.061	-0.959	0.043	0.061
	.80	-6.933	0.047	0.059	6.284	0.048	0.059	0.498	0.048	0.059
	.90	-10.010	0.048	0.054	6.893	0.049	0.054	-0.551	0.049	0.054
	.10	-0.739	0.033	0.336	1.036	0.034	0.336	0.266	0.034	0.336
	.20	-2.974	0.035	0.176	0.405	0.035	0.176	-1.063	0.035	0.176
	.30	-1.773	0.035	0.117	2.985	0.035	0.117	0.921	0.035	0.117
	.40	-5.661	0.034	0.086	0.576	0.035	0.086	-2.138	0.034	0.086
40	.50	-3.755	0.034	0.069	4.360	0.035	0.069	0.832	0.035	0.069
	.60	-4.938	0.036	0.060	6.439	0.036	0.060	1.462	0.036	0.060
	.70	-5.189	0.035	0.050	6.049	0.036	0.050	1.137	0.035	0.050
	.80	-5.609	0.036	0.045	8.407	0.037	0.045	2.260	0.036	0.045
	.90	-7.732	0.036	0.041	8.185	0.037	0.041	1.183	0.037	0.041
	.10	-1.099	0.026	0.265	0.653	0.027	0.265	-0.107	0.026	0.265
	.20	-0.485	0.027	0.134	2.692	0.027	0.134	1.315	0.027	0.134
	.30	-2.958	0.027	0.091	1.743	0.028	0.091	-0.296	0.027	0.091
	.40	-3.897	0.027	0.068	3.241	0.027	0.068	0.134	0.027	0.068
67	.50	-2.544	0.028	0.057	6.147	0.029	0.057	2.364	0.029	0.057
	.60	-4.538	0.027	0.046	5.583	0.028	0.046	1.164	0.028	0.046
	.70	-5.724	0.027	0.039	6.290	0.027	0.039	1.035	0.027	0.039
	.80	-7.448	0.028	0.035	5.632	0.028	0.035	-0.103	0.028	0.035
	.90	-7.815	0.030	0.033	8.299	0.030	0.033	1.217	0.030	0.033

Table 5

Bias, Standard Deviation, and Coefficient of Variation of $\hat{\delta}_{WLS}$, $\hat{\delta}_{ML}^{(old)}$ and $\hat{\delta}_{ML}^{(new)}$

(Average sample sizes $n_e = 100$, $n_c = 100$)

		$\hat{\delta}_{WLS}$			$\hat{\delta}_{ML}^{(old)}$			$\hat{\delta}_{ML}^{(new)}$		
k	δ	bias$\times 10^3$	SD	CV	bias$\times 10^3$	SD	CV	bias$\times 10^3$	SD	CV
	.10	0.485	0.032	0.314	1.314	0.032	0.314	0.957	0.032	0.314
	.20	-0.772	0.033	0.166	1.183	0.033	0.166	0.338	0.033	0.166
	.30	-0.714	0.032	0.107	1.700	0.032	0.107	0.658	0.032	0.107
	.40	-2.601	0.035	0.088	0.657	0.035	0.088	-0.751	0.035	0.088
24	.50	-4.004	0.030	0.061	-0.294	0.031	0.061	-1.899	0.030	0.061
	.60	-3.960	0.033	0.056	1.026	0.034	0.056	-1.144	0.033	0.056
	.70	-4.703	0.029	0.042	0.286	0.030	0.042	-1.882	0.030	0.042
	.80	-3.666	0.033	0.041	2.854	0.033	0.041	0.013	0.033	0.041
	.90	-5.976	0.033	0.037	2.425	0.034	0.037	-1.254	0.033	0.037
	.10	-0.378	0.024	0.243	0.501	0.024	0.243	0.122	0.024	0.243
	.20	-0.320	0.025	0.127	1.362	0.026	0.127	0.636	0.026	0.127
	.30	-0.491	0.025	0.085	1.842	0.026	0.085	0.835	0.026	0.085
	.40	-1.231	0.024	0.059	1.848	0.024	0.059	0.515	0.024	0.059
40	.50	-2.787	0.026	0.053	1.212	0.027	0.053	-0.518	0.026	0.053
	.60	-2.499	0.026	0.044	3.127	0.026	0.044	0.682	0.026	0.044
	.70	-3.124	0.023	0.033	2.419	0.023	0.033	0.008	0.023	0.033
	.80	-4.514	0.025	0.032	2.398	0.026	0.032	-0.617	0.026	0.032
	.90	-4.587	0.026	0.029	3.248	0.027	0.029	-0.179	0.026	0.029
	.10	-0.765	0.020	0.201	0.116	0.020	0.201	-0.264	0.020	0.201
	.20	0.006	0.019	0.096	1.563	0.019	0.096	0.892	0.019	0.096
	.30	-0.931	0.019	0.062	1.389	0.019	0.062	0.387	0.019	0.062
	.40	-0.807	0.018	0.046	2.744	0.019	0.046	1.207	0.018	0.046
67	.50	-1.640	0.019	0.038	2.644	0.019	0.038	0.789	0.019	0.038
	.60	-1.796	0.019	0.032	3.200	0.019	0.032	1.031	0.019	0.032
	.70	-2.715	0.019	0.027	3.231	0.019	0.027	0.643	0.019	0.027
	.80	-4.396	0.018	0.022	2.018	0.018	0.022	-0.781	0.018	0.022
	.90	-3.030	0.020	0.022	4.937	0.020	0.022	1.453	0.020	0.022

Table 6

Number of Iterations for EM-algorithm to Converge in Four Empirical Data Sets

		WLS Estimator		Old MLE		New MLE	
Sources[1]	k	$\hat{\delta}_{WLS}$	No. of Iterations	$\hat{\delta}_{ML}^{(old)}$	No. of Iterations	$\hat{\delta}_{ML}^{(new)}$	No. of Iterations
Table 6	11	.286 (.056)	N/A	.290 (.055)	3	.288 (.056)	5
Table 7	7	-.097 (.099)	N/A	-.100 (.098)	2	-.098 (.099)	5
Table 8	18	.011 (.042)	N/A	.010 (.042)	2	.011 (.042)	4
Table 9	10	.054 (.080)	N/A	.061 (.078)	3	.058 (.080)	6

[1]Based on the *effects of open education* as reported in Tables 6 to 9 of Hedges and Olkin (1985, pp. 23–25).

Table 7

The Effect of Extreme Effect Size and Sampe Size on $\hat{\delta}_{WLS}$, $\hat{\delta}_{ML}^{(old)}$ and $\hat{\delta}_{ML}^{(new)}$

Based on Effects of Open Education Data from Table 6 of Hedges and Olkin (1985, p. 24)

	$\hat{\delta}_{WLS}$		$\hat{\delta}_{ML}^{(old)}$		$\hat{\delta}_{ML}^{(new)}$	
Data	$\hat{\delta}_{WLS}$	No. of Iterations	$\hat{\delta}_{ML_O}$	No. of Iterations	$\hat{\delta}_{ML_N}$	No. of Iterations
Original Data	.286	N/A	.290	1	.288	5
$(n_{e12} = n_{c12} = 2)$						
$d_{12} =$						
1	.287	N/A	.293	3	.290	5
2	.288	N/A	.294	3	.292	5
3	.289	N/A	.295	3	.292	5
4	.290	N/A	.295	3	.292	5
5	.290	N/A	.295	3	.292	5
-1	.284	N/A	.286	2	.285	5
-2	.282	N/A	.284	2	.284	5
-3	.282	N/A	.284	2	.283	5
-4	.281	N/A	.284	2	.283	5
-5	.281	N/A	.284	2	.283	5
$d_{12} = 3$						
$n_{e12} = n_{c12} =$						
5	.296	N/A	.302	3	.299	5
10	.306	N/A	.313	3	.310	5
20	.325	N/A	.335	3	.332	6
40	.363	N/A	.378	3	.376	6
80	.435	N/A	.460	3	.457	7
$d_{12} = -3$						
$n_{e12} = n_{c12} =$						
5	.274	N/A	.275	2	.275	5
10	.262	N/A	.261	2	.261	5
20	.239	N/A	.233	3	.233	6
40	.193	N/A	.180	3	.179	6
80	.106	N/A	.078	3	.077	7

Figure 1. Likelihood function of noncentral t-distribution with degrees of freedom 8, observed $x = (5/2)^{1/2}$ and noncentrality parameter $(5/2)^{1/2}\delta$

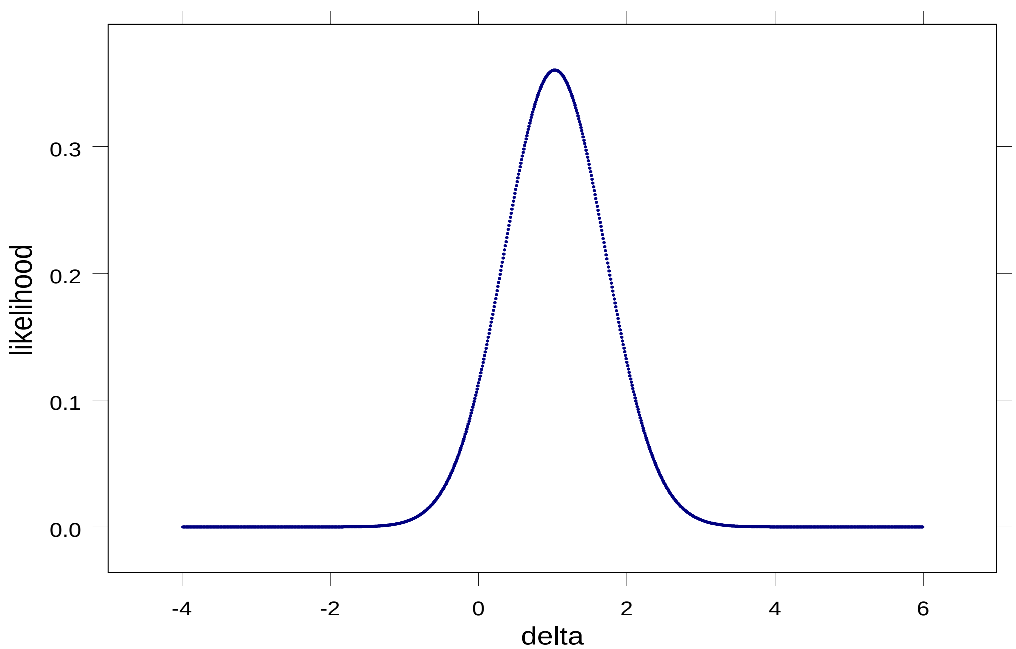
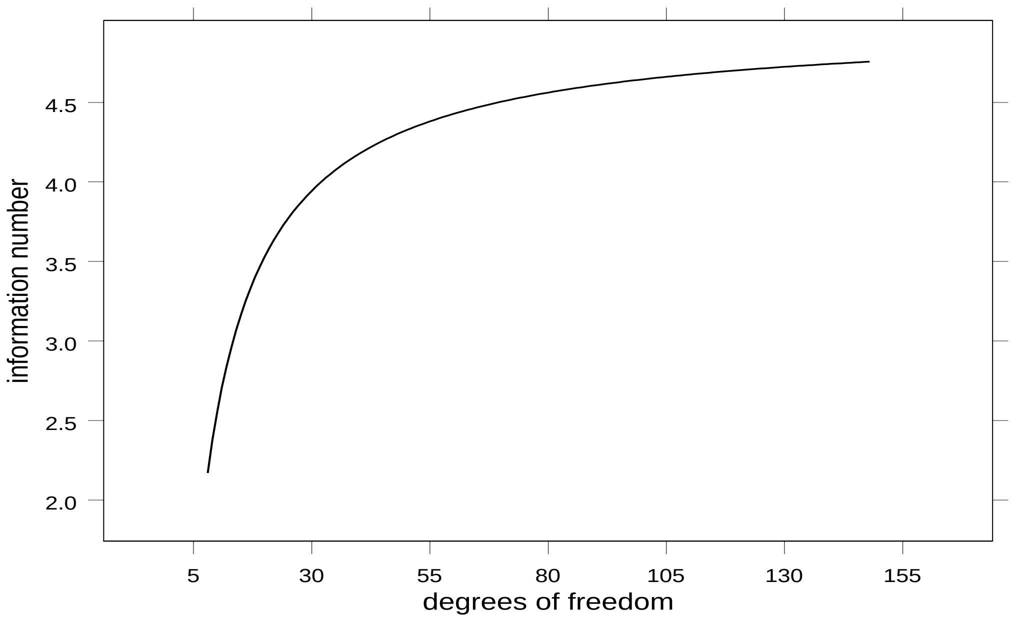


Figure 2. Information versus degrees of freedom

(a) $n_e = 5$, $n_c = 5 - -145$, $d = 1.0$



(b) $n_e = n_c = 5 - -75$, $d = 1$

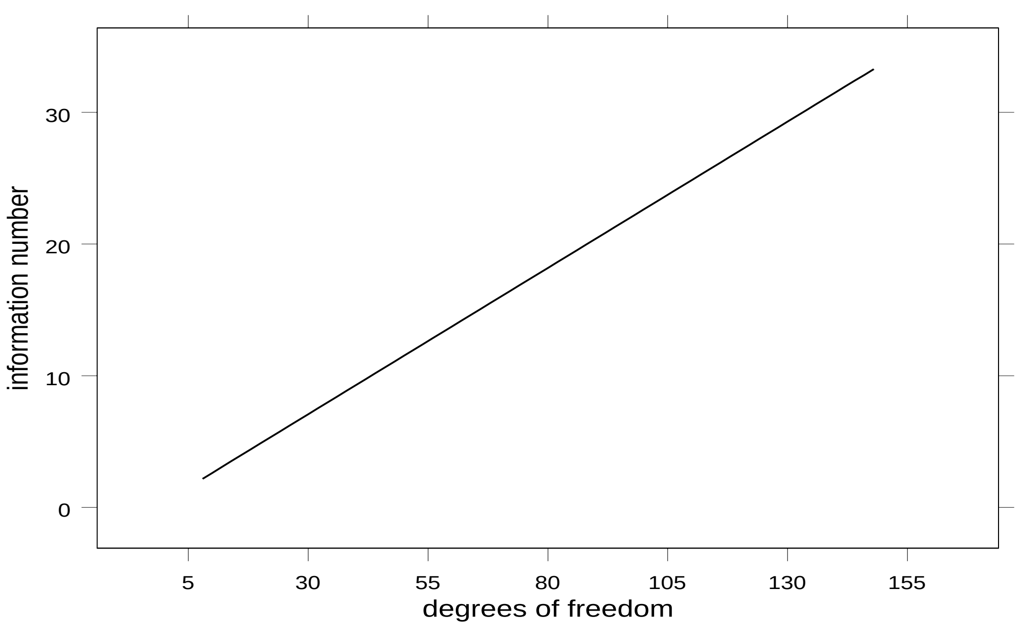
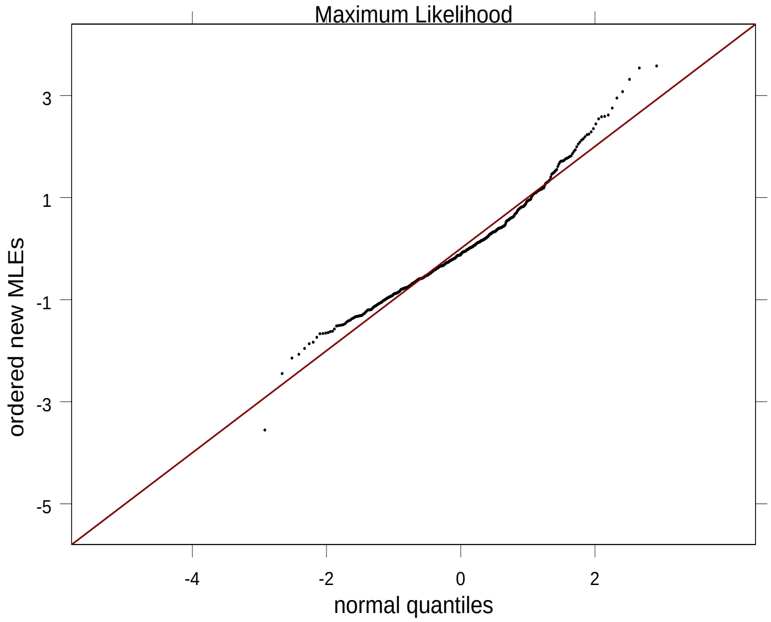
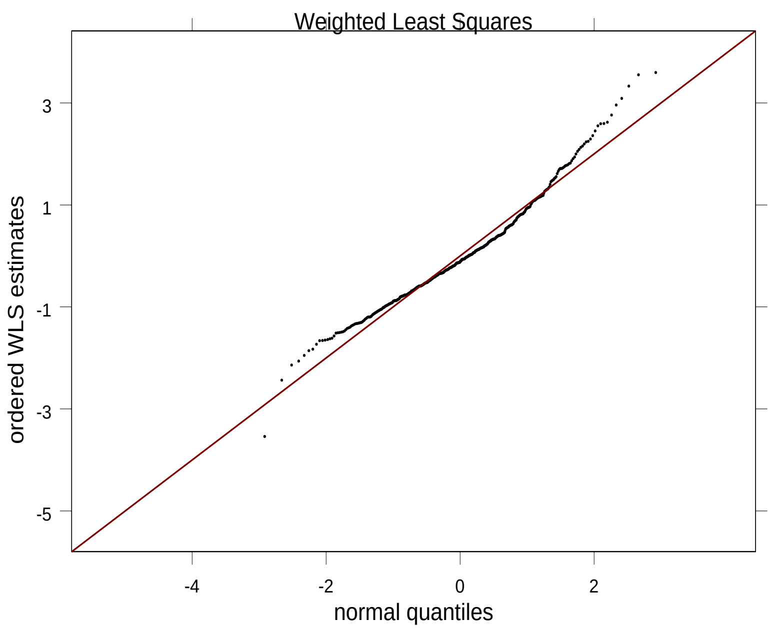


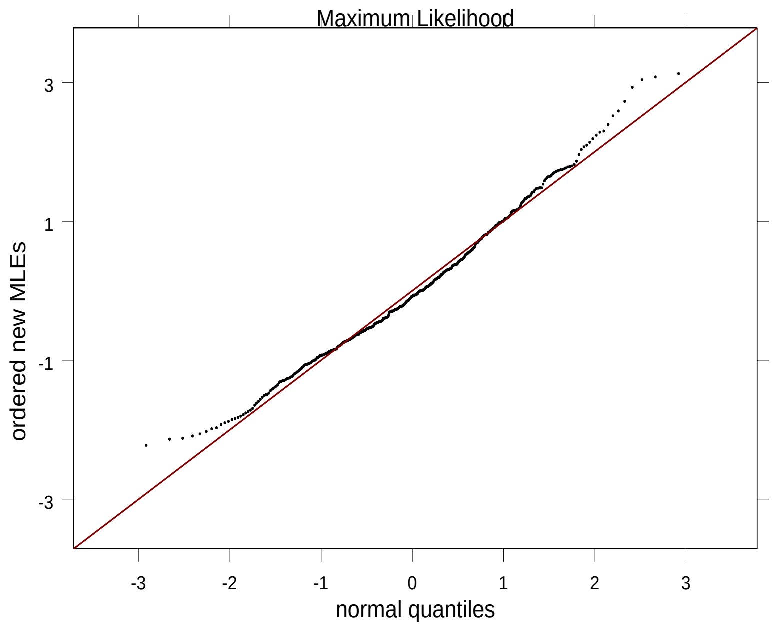
Figure 3. Normal probability plots for $\hat{\delta}_{ML}^{(new)}$ and $\hat{\delta}_{WLS}$ based on a single effect size with 500 replications

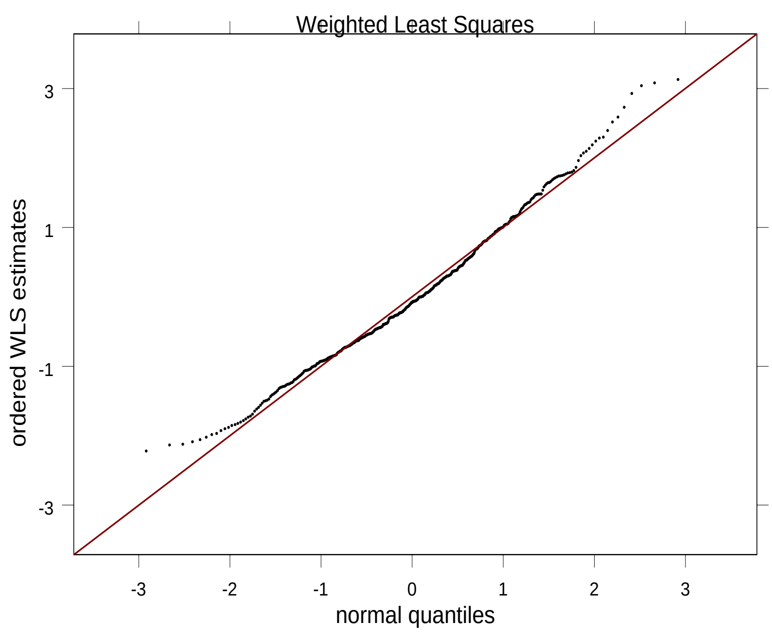
(a) $n_e = 10$, $n_c = 10$, $\delta = 1$



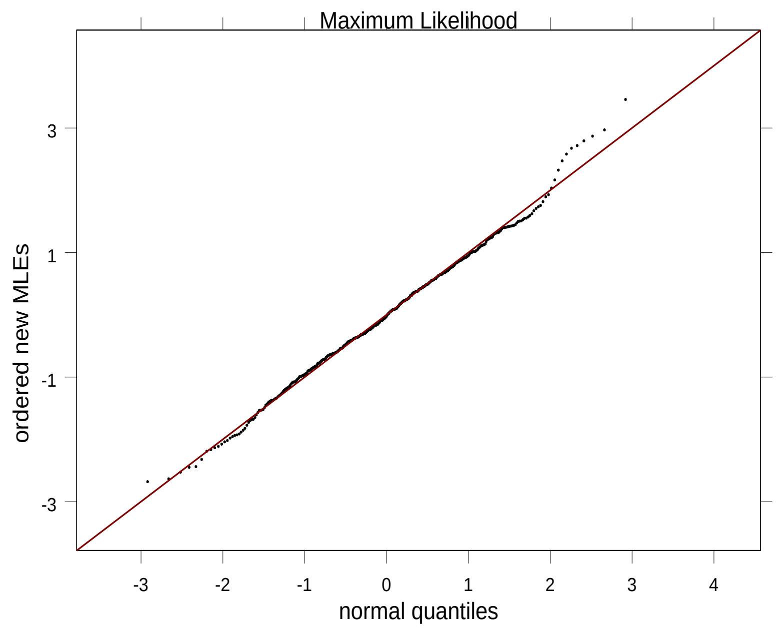


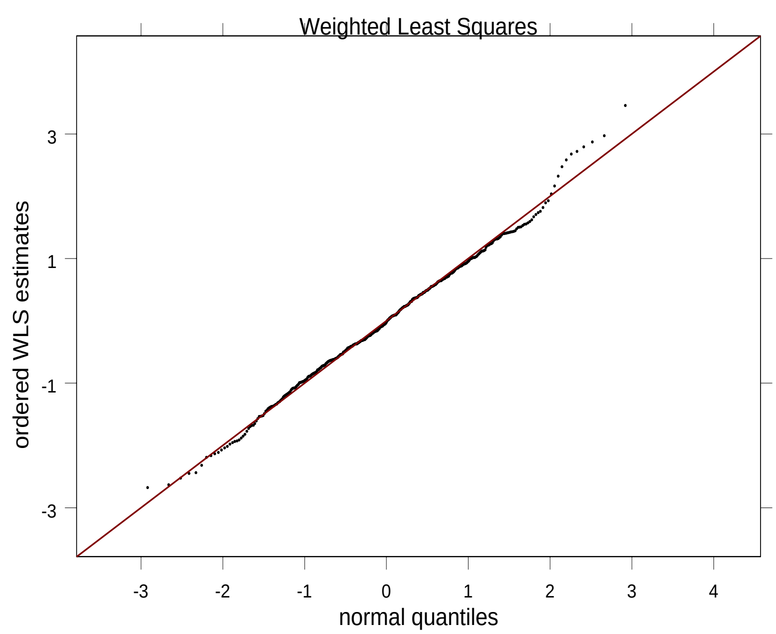
(b) $n_e = 20$, $n_c = 20$, $\delta = 1$





(c) $n_e = 40$, $n_c = 40$, $\delta = 1$





(d) $n_e = 80$, $n_c = 80$, $\delta = 1$