ABSTRACT

Dissertation Title: FINITE MIXTURE MODEL SPECIFICATIONS
ACCOMMODATING TREATMENT NONRESPONSE IN
EXPERIMENTAL RESEARCH

John Andrew Wasko, Doctor of Philosophy, 2009

Directed by: Professor Gregory R. Hancock, Department of Measurement,
Statistics and Evaluation

For researchers exploring causal inferences with simple two group experimental
designs, results are confounded when using common statistical methods and further are
unsuitable in cases of treatment nonresponse. In signal processing, researchers have
successfully extracted multiple signals from data streams with Gaussian mixture models,
where their use is well matched to accommodate researchers in this predicament. While
the mathematics underpinning models in either application remains unchanged, there are
stark differences. In signal processing, results are definitively evaluated assessing
whether extracted signals are interpretable. Such obvious feedback is unavailable to
researchers seeking causal inference who instead rely on empirical evidence from
inferential statements regarding mean differences, as done in analysis of variance
(ANOVA). Two group experimental designs do provide added benefit by anchoring
treatment nonrespondents' distributional response properties from the control group.

Obtaining empirical evidence supporting treatment nonresponse, however, can be extremely challenging. First, if indeed nonresponse exists, then basic population means, ANOVA or repeated measures tests cannot be used because of a violation of the identical distribution property required for each method. Secondly, the mixing parameter or proportion of nonresponse is bounded between 0 and 1, so does not subscribe to normal distribution theory to enable inference by common methods.

This dissertation introduces and evaluates the performance of an information-based methodology as a more extensible and informative alternative to statistical tests of population means while addressing treatment nonresponse. Gaussian distributions are not required under this methodology which simultaneously provides empirical evidence through model selection regarding treatment nonresponse, equality of population means, and equality of variance hypotheses. The use of information criteria measures as an omnibus assessment of a set of mixture and non-mixture models within a maximum likelihood framework eliminates the need for a Newton-Pearson framework of probabilistic inferences on individual parameter estimates. This dissertation assesses performance in recapturing population conditions for hypotheses' conclusions, parameter accuracy, and class membership. More complex extensions addressing multiple treatments, multiple responses within a treatment, *a priori* consideration of covariates, and multivariate responses within a latent framework are also introduced.

FINITE MIXTURE MODEL SPECIFICATIONS ACCOMMODATING TREATMENT
NONRESPONSE IN EXPERIMENTAL RESEARCH

By

John Andrew Wasko

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2009

Advisory Committee:
Professor Gregory R. Hancock, Chair
Professor Robert Lissitz
Professor Robert Mislevy
Professor Robert Croninger
Professor Jeffrey Harring

Acknowledgements

When the requirement for obtaining a doctoral degree is advancement or innovation in a chosen field of study, an inherent consequence is an often lonely and difficult journey into unchartered territory. Having survived the experience with relatively sound mind and body, there are some whom I would like to extend my most sincere thanks.

To the military, who offered this opportunity. I applied. I was accepted. Having completed my studies and returning to more traditional military duties, my hope is you are receiving a more capable soldier. With 22 years of service now, it is safe to say I plan on making a career in the Army.

To my family, who understood and/or cared very little about my dissertation topic and wholly about my well being. Their love and support made the journey less lonely and their patience and understanding of many late night and middle of the night trips to the basement cave (office) allowed me to complete my task in a reasonable period of time.

To my advisor, Greg Hancock, who cares equally about the topic and the student. He has many titles and responsibilities: Chair, Professor, Doctor, husband and father. And while I have observed him in each of these capacities, I am most honored in referring to him by yet another title, friend.

In many ways, I feel like the scarecrow from the Wizard of Oz, who received the following commentary from the wizard.

> *"Why, anybody can have a brain. That's a very mediocre*
> *commodity. Every pusillanimous creature that crawls on*
> *the Earth or slinks through slimy seas has a brain. Back*

Acknowledgements (continued)

> *where I come from, we have universities, seats of great learning, where men go to become great thinkers. And when they come out, they think deep thoughts and with no more brains than you have. But they have one thing you haven't got: a diploma."*

University of Maryland is indeed a great seat of learning, and while no one would ever confuse me with a great thinker, this mediocre commodity can think deep thoughts and has earned his hopefully last diploma.

Table of Contents

Table of Contents (continued)

List of Tables

List of Figures

List of Figures (continued)

List of Figures (continued)

Chapter 1: Introduction

The area of finite mixture models is sufficiently large and well established as a viable method useful in a number of disciplines for explaining relationships in observed data. There are a number of general excellent resources including Everitt and Hand (1981) and McLachlan and Peel (2000). Medical research such as Boos (1991) and Luo (2004) applied mixtures in a regression based framework more broadly described for generalized linear models (GLM) by Wedel and Desarbo (1995). Others, such as Pavlic (2001), considered mixtures in two group experimental designs with applied research and simulation studies. The nature of these applications and empirical studies are varied; some simulations illustrated the viability of a mixture model representation, others specified elaborate models representative of a particular data set, while others focused on a posterior probability aspect of individual group membership. None of these articles, however, endeavored to provide empirical evidence supporting the theoretical supposition of treatment nonresponse. Treatment nonresponse is *not* that a respondent provides no measurement; it is an individual, in full compliance with the particular treatment, demonstrates no change in measured response from the treatment compared with the baseline group. The approach advocated in this dissertation represents an innovative solution strategy combining a number of current analytic techniques taking advantage of sound experimental design. Further, this strategy serves as a suitable foundation upon which to evaluate more complex research questions in a number of different directions presented in Chapter 4.

A common experimental design is a two group posttest-only randomized

experimental scenario consisting of a control and treatment, or even competing

treatments. Many extensions are available from this basic design: inclusion of a pretest,

multiple groups, factorial conditions facilitating systematic manipulation of several

independent variables, and randomized block designs are a few examples. When

randomization is neither practical nor feasible, quasi-experimental variations such as the

non-equivalent groups' pretest-posttest design are available. Unfortunately, when

implementing any of these experimental designs, the effectiveness of any inference

becomes unclear under consideration of treatment nonresponse. A commonly held

assumption is that individuals respond representative of a homogeneous population,

albeit with varied responses. The subsequent goal is to estimate the impact of the

treatment in comparison to the control. In reality, however, some treatments might only

yield results for some members while failing completely for others within the same

sample. In such cases, treatment sample members represent two populations: treatment

responders, the basis for any causal inferences, and treatment nonresponders, acting act

as a contaminant, where unfortunately membership is unknown. In medical applications,

explanations of treatment nonresponse to medications are commonly attributed to some

type of physiological phenomenon. In the social sciences, however, the definition of a

treatment has a broader application including additional instruction, supplemental

training, changes in environmental settings, or even material presented under a different

pedagogy. Without such tangible physiological explanations, researchers attribute

nonresponse to these particular treatments as a simple matter of compliance by an

individual. Just as likely is the possibility, having fully complied with the treatment, that

no change in the cognitive processes or behavior occurred in these individuals assessed in

a post treatment evaluation.

The difficulty in this seemingly straightforward problem is assessing and mitigating the contamination in the treatment group. A foundational assumption of IID, *independence* and *identical distribution*, exists in all population means tests, whether a particular *t*-test or in ANOVA. In certain cases, one is confident that the "*I*", independence of observations, is violated. Repeated measures in the paired *t*-test, longitudinal data analysis, hierarchical linear modeling, and time series analysis (AR, ARIMA, etc.) parametrically account for this violation. Upon mitigation, researchers return to the comfort of IID upon which subsequent parameter and causal inferences are made. For each of these modeling techniques or statistical tests, the assumption of *"ID"*, identically distributed observations, persists. Treatment nonresponse, if it exists, is a violation of the ID assumption where respondents can no longer be aggregated into a homogeneous group.

There are currently three choices if one supposes treatment nonresponse exists in a particular sample. The most mathematically convenient and easiest choice is to ignore it, which allows the researcher to continue with traditional analysis presented in the next section. The consequence of such a decision is the degree to which inferences regarding a *true* treatment effect are clouded, where the estimate of the treatment population mean is affected beyond sampling error. The typical Type I error control, α, used as a threshold in declaring statistical significance, becomes unreliable. On one hand, the clouding may be so minute, for a number of reasons subsequently addressed, that the overall conclusion regarding a difference in population means is unchanged. Of greater concern is the possibility that one falsely concludes no difference in the

population means. In either case, the clouding is such the expected estimate of the

population mean for the treatment is negatively biased due to treatment nonresponse.

Ignoring nonresponse and aggregating as a homogeneous group, treatment effects would

be underestimated by up to 70% in the range of experimental conditions subsequently

evaluated (mathematical expressions for these expected biases are presented in section

2.2.1).

A second option, much more mathematically complex, addresses the lack of

homogeneity or contamination as a mixture of two components: responders and non-

responders. Such models introduce a mixing proportion parameter, ϕ, in model

specification, usually determined within a maximum likelihood framework. Both choices

conduct analysis, though neither option has any empirical evidence regarding

nonresponse supporting their model specification or particular statistical test. Using the

first option is tantamount to concluding a homogeneous treatment response, while the

latter is also implicitly drawing a conclusion: nonresponders are present.

A final option, a combination of the previous two, incorporates the concept of

treatment nonresponse with the expediency available in traditional analysis include Intent

to Treat (ITT), and Complier Average Casual Effect (CACE) models. CACE models

introduce a measured variable associated with a degree of an individual's treatment

compliance as a polytomous indicator, where ANOVA or regression can be subsequently

utilized as presented in Angrist, Imbrens, and Rubin (1996). This dissertation assumes a

single treatment with uniform compliance among sample respondents. ITT might be

considered a viable alternative if a researcher aggregates individual results to perform

analysis on a particular treatment. Hollis (1999) and Lachin (2000) provided a cursory

look at this particular strategy principally employed to address randomization challenges in clinical trials. Comparison of surgical versus drug therapy options for a medical condition is an example, where individuals are not *assigned* and measured outcomes may take months or years post treatment to assess. ITT is flexible to allow degrees of compliance, like CACE models, and particular to this study, can and will prescribe different treatments based on an assessed ineffectiveness of the current treatment. While ITT addresses the issue of randomization supporting experimental designs, it is not a defensible strategy supporting population treatment nonresponse. At some point, ITT might transition from focus on the individual to conduct group level analysis for population inferences after individual classification for a particular treatment. Using the post-treatment response, and perhaps other covariate information, an individual is classified as a respondent or nonrespondent. The overall mixing proportion estimate for the population comes from the summation of these individual assessments. Upon classification, traditional analysis methods become available as this established two *ID* groups based wholly, or in part, on the post-treatment response.

Used in this manner, this option suffers from two drawbacks, including false classification and a lack of probabilities for group membership, both related to individual assessment. The classification process is assumed error free, in other words, there are no false classifications, either a nonrespondent as a respondent or respondent as a nonrespondent. False classifications are analogous to non-*ID* groups which in turn affect the conclusions of any subsequent analysis. However unlikely error free classifications may be to a particular treatment with varied responses, one would expect some non-zero misclassification probability across the sample of treatment respondents.

A more reasonable modeling approach results in a probability or likelihood to each individual regarding treatment nonresponse instead of a dichotomous assignment. Drawing inferences about an individual's probability of being in either of these possible classes using their measured response is an application of Bayes' theorem. Often referred to as posterior probabilities, these are more formally presented in section 2.3.3. Obtaining these probabilities, unfortunately, is not tractable when focused at the individual level because it lacks population distribution properties to complete the calculations. Subsequently, ITT fails to address, even from a probability perspective, empirical evidence regarding treatment nonresponse. Using a mixture framework, however, where unknown nonresponders have the same distributional properties in their response as the control population, one can obtain a mixing proportion estimate in lieu of summarizing individual classifications. A mixing proportion estimate obtained in this fashion in turn enables transition from the population to each individual in the sample with calculation of posterior probabilities of treatment nonresponse membership.

Consistent among all three options is the lack of empirical evidence as part of model selection regarding nonresponse. Each option has made an *a priori* decision with regard to an analysis technique and ID assumption. A supposition of treatment nonresponse implies some sort of statistical assessment is in order to substantiate this claim, where a comparative model framework utilizing information criteria for model selection can serve as the basis for statistical assessment. This dissertation proposes a solution beyond individual parameter assessment to a broader level of comparative model evaluation as the means of obtaining empirical evidence simultaneously for the following hypotheses:

6

$$Ho_1 : \phi = 0 \text{ - Homogeneous } (ID) \text{ population}$$
$$Ha_1 : \phi > 0 \text{ - Treatment Nonresponse Exists}$$

$$Ho_2 : \mu_C = \mu_T \text{ - No Treatment Effect}$$
$$Ha_2 : \mu_C \neq \mu_T \text{ - Treatment Effect}$$

When a normal distribution is posited for the control or treatment population, an additional hypothesis may be assessed concurrently:

$$Ho_3 : \sigma_C^2 = \sigma_T^2 \text{ - Equal Population Variances}$$
$$Ha_3 : \sigma_C^2 \neq \sigma_T^2 \text{ - Unequal Population Variances}$$

In model representations positing a homogeneous, identically distributed response, a holistic model selection utilizing information criteria serves as an alternative to common statistical tests on population means and variances. There is no analog when introducing models including treatment nonresponse, yet holistic model selection readily accommodates this condition in a broader set of models.

The central question of the existing three options and the proposed methodology is the same: hypothesis statements regarding differences in population means. Unlike the other options, however, this methodology additionally provides empirical evidence regarding treatment nonresponse, both in a manner which does not require normal theory probabilistic inferences on individual parameter estimates.

1.1 Two Group Experimental Designs and Traditional Analysis

Two group treatment/control designs serve as the workhorse in research studies, often used as the launching point into complex multi-factorial designs, multivariate responses, or more involved research interests. The methodological development and empirical evaluation in this dissertation focuses on the most basic design. Experimental

design considerations such as a pseudo-random or specific selection study, assessment

methods, types of measurement, and controlled conditions must be defended by the

researcher.

Figure 1 sets the backdrop for these experimental designs within a broader

family of procedures used to make inferences regarding population means from samples

(Park, 2008). These designs fall within the shaded region where, without considering

treatment nonresponse, corresponding statistical tests are readily available in SAS, Stata,

SPSS, and other statistical software. This section is intended only as a re-familiarization

as a building block to an alternative approach and subsequent consideration of

nonresponse. Kirk (1995) and Keppel (2004) offered more detailed formulation for each

of these test procedures.

Figure 1.
Schematic for Population Means Test Procedures with No Mixture
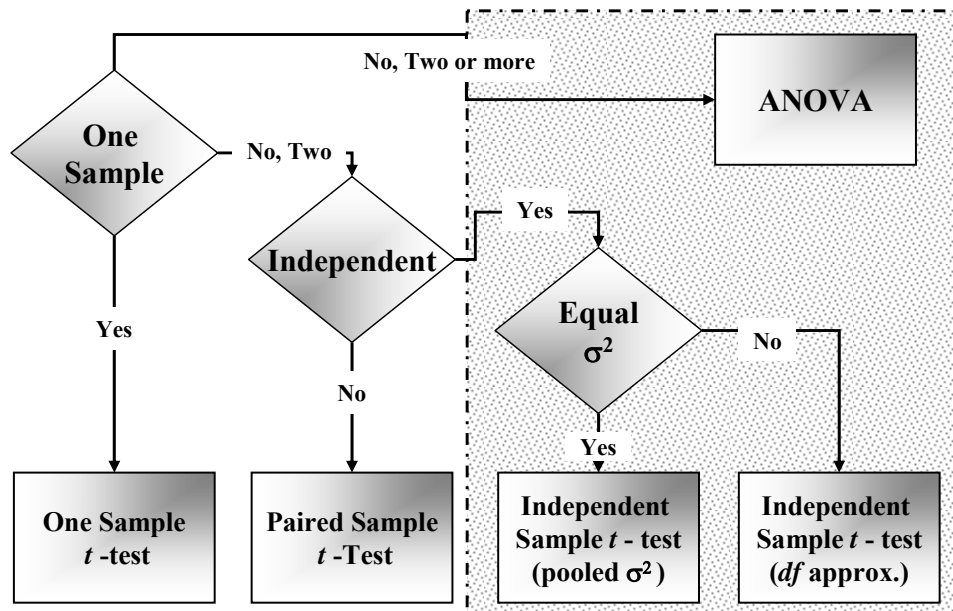
The shaded region in the figure highlights traditional two sample statistical

procedures also indicating independence required between samples. Another required

assumption for these tests is identically distributed observations, which corresponds, for

example to an absence of treatment nonresponse or mixture in the treatment sample. Let

$$\mathbf{X}_{Ci} = x_{C1}, x_{C2}, ..., x_{Cn_c} \sim f_C(\boldsymbol{\theta}) \quad \text{for } i = 1, 2, ..., n_C \tag{1}$$

represent responses for a control sample of size n_C which can be characterized by some

parametric distribution. For a treatment sample, assuming <u>no</u> mixture in the sample,

then

$$\mathbf{X}_{Tj} = x_{T1}, x_{T2}, ..., x_{Tn_T} \sim f_T(\boldsymbol{\theta}) \quad \text{for } j = 1, 2, ..., n_T \tag{2}$$

represents the measured responses for a treatment sample of size n_T characterized by

some parametric distribution. It is commonly assumed, but not necessary, that

$$f_C(\boldsymbol{\theta}) = \text{Nor}(\mu_C, \sigma_C)$$
$$f_T(\boldsymbol{\theta}) = \text{Nor}(\mu_T, \sigma_T) \tag{3}$$

An assumption of normality is, in most cases, quite reasonable. Preference for the

normal distribution is due to its unique flexibility of distributional parameter

independence for its first two moments, location and scale, facilitating excellent

characterization of many types of observed responses. For sufficiently large sample sizes,

n_C and n_T, the distribution of the mean is well represented by a normal distribution

regardless of the population distribution under the protection of the central limit theorem

where

$$\bar{x}_C \sim \text{Nor}\left(\mu_C, \frac{\sigma_C}{\sqrt{n_C}}\right) \tag{4}$$

$$\bar{x}_T \sim \text{Nor}\left(\mu_T, \frac{\sigma_T}{\sqrt{n_T}}\right) \tag{5}$$

with $\bar{x}$ and s used as estimates for the respective population parameters. Because of this

result, both statistical tests yield consistent results if one or both samples display skewed

or kurtotic properties. If the researcher believes the population variances are equal, then

a pooled variance t - test statistic can be computed where

$$t^* = \frac{\bar{x}_T - \bar{x}_C}{\sqrt{s_P^2\left(\frac{1}{n_C} + \frac{1}{n_T}\right)}} \tag{6}$$

and

$$s_P^2 = \frac{(n_C - 1)s_C^2 + (n_T - 1)s_T^2}{n_C + n_T - 2} \tag{7}$$

which subscribes to a central t distribution with $n_C + n_T - 2$ degrees of freedom (df).

However, knowledge of population variances cannot be achieved without knowledge of

population means. Moser (1992) accordingly recommended prior to its use a preliminary

F- test assessing variance equality. Yet Gans (1984) noted the combination of two

statistical tests is problematic in controlling Type I error for population means inferences

with the accompanying recommendation to always use the unequal variance t-test,

commonly attributed to Satterwaite (1946) and Welch (1938). This test statistic is

calculated as

$$t^* = \frac{\bar{x}_T - \bar{x}_C}{\sqrt{\frac{s_C^2}{n_C} + \frac{s_T^2}{n_T}}} \tag{8}$$

and also follows a central t distribution where the degrees of freedom, v, is approximated

$$v \approx \frac{\left(\frac{1}{n_C} + \frac{s_T^2/s_C^2}{n_T}\right)^2}{\frac{1}{n_C^2(n_C - 1)} + \frac{(s_T^2/s_C^2)^2}{n_T^2(n_T - 1)}} \tag{9}$$

Either test statistic is compared to a $(1-\alpha) \times 100\%$ critical value from the central t distribution with the respective df, where larger df results in probability values approaching a standard normal distribution. In the context of this problem, a researcher would conduct a one-sided hypothesis test dependent on the theory whether the treatment comparatively raises or lowers the measured response. The result is a reject or a failure to reject the null hypothesis of population means equality, where a rejection indicates a treatment effect. Preference for the use of the Satterwaite t - test is further bolstered by Coombs (1996) and Zimmerman (2004) who showed the robustness of this statistic against differences in samples sizes and population variances, unlike the pooled variance test.

Two sample experimental designs can also accommodate repeated measures. Consider two independent groups, each receiving pre and post-treatment measurements, where the control group is administered a placebo treatment. Setting $t = 0$ as the baseline measurement and $t = 1$ as the post-treatment measurement, data of this form can be represented akin to Equations (1) and (2) where

$$X_{Ci} = (x_{t=1,C1} - x_{t=0,C1}) = (x_{1C1} - x_{0C1}),...,(x_{1Cn_c} - x_{0Cn_c}) \tag{10}$$

and

$$X_{Tj} = (x_{t=1,T1} - x_{t=0,T1}) = (x_{1T1} - x_{0T1}),...,(x_{1Tn_T} - x_{0Tn_T}) \tag{11}$$

subsequently using either of the t - tests presented.

For multiple subgroups, analysis of variance (ANOVA) can be utilized to include two sample comparisons as previously discussed. This is slightly different from the t - tests where random effects ANOVA considers these samples as two sub-groups

from a single population, with a more stringent requirement of normality in the population and sub-groups in addition to variance equality.

1.2 An Information Criterion Analog

Considering the experimental designs presented, a researcher can also obtain empirical evidence of a treatment effect through a series of comparative model assessments. Instead of relying on summary statistics, $\bar{x}$ and s^2, for calculations and probabilistic inferences, this can be viewed in a larger context: fitting population distribution parameters to a series of models of increasing parsimony through parameter equality constraints. Both methods provide empirical evidence, yet this alternative is information based rather than supported by probabilistic inferences. While more computationally intensive than the traditional methods presented, it offers added benefits. First, it removes inferential statements regarding population parameters where hypothesis conclusions are based on arbitrary Type I error control thresholds, be it .01, .05, or even .10. Second, it simultaneously provides empirical evidence regarding variance equality, not evaluated in the Satterwaite t - test and is a separate, preliminary statistical test with the pooled variance t - test. Evidence of differing variability for a treatment, producing a more consistent or widespread population response, provides a depth of treatment information beyond the standard reporting afforded with current traditional testing.

Estimation of population distribution parameters can be accomplished within a maximum likelihood (ML) framework. Maximum likelihood is a mathematical optimization process resulting in the most likely set of parameters for a given model specification using the data. A general form of the two sample representation is

$$L(\mathbf{X}\,|\,\boldsymbol{\theta}) = L(\mathbf{X}_C\,|\,\boldsymbol{\theta}_C)*L(\mathbf{X}_T\,|\,\boldsymbol{\theta}_T) = \prod_{i=1}^{n_c} f_C(x_{Ci}\,|\,\boldsymbol{\theta}_C)*\prod_{j=1}^{n_T} f_T(x_{Tj}\,|\,\boldsymbol{\theta}_T) \qquad (12)$$

where $\boldsymbol{\theta}$ represents the vector of population distributional parameters, $\boldsymbol{\theta}_C$ and $\boldsymbol{\theta}_T$. This

complicated formula involves the product of products making computations difficult. To

simplify, vertically concatenate the control and treatment sample observations such that

$$\mathbf{Y}_N = x_{C1}, x_{C2},..., x_{Cn_c}, x_{Tn_c+1}, x_{Tn_c+2},..., x_{Tn_c+n_T} \qquad N = n_C + n_T \qquad (13)$$

and introduce a new term of the form

$$h_i = \begin{cases} 1 & \text{if control} \qquad \text{for } i = 1,2,.., N \\ 0 & \text{if treatment} \end{cases} \qquad (14)$$

where Equation (12) can be rewritten as

$$L(\mathbf{Y}\,|\,\boldsymbol{\theta}) = \prod_{i=1}^{N}\left[h_i f_C(y_i\,|\,\boldsymbol{\theta}_C) + (1-h_i)f_T(y_i\,|\,\boldsymbol{\theta}_T)\right] \qquad (15)$$

This step reduced the likelihood function to a single product term, yet remains

computationally challenging because each term results in a probability density function

(PDF) or probability mass function (PMF) value ranging between [0,1]. The product of

positive fractional numbers creates an extremely small positive number, affecting

precision and degrading the ability of search algorithms to maximize this function. To

mitigate these problems, a logarithmic transform can be performed without changing the

location of the maximum value of $L(\mathbf{Y}\,|\,\boldsymbol{\theta})$. Therefore, the general equation used to

determine parameter estimates becomes

$$\ln\!\left(L(\mathbf{Y}\,|\,\boldsymbol{\theta})\right) = \sum_{i=1}^{N}\ln\!\left(h_i f_C(y_i\,|\,\boldsymbol{\theta}_C) + (1-h_i)f_T(y_i\,|\,\boldsymbol{\theta}_T)\right) \qquad (16)$$

Further development will use, as an example, normal distributions for the

control and treatment populations. This is not a requirement for either population;

nonetheless, the normal distribution remains a common and flexible parametric representation. The PDF of the normal distribution for a single observation x is

$$f(x \mid \boldsymbol{\theta}) = f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2}\right) \tag{17}$$

where population parameter estimates for a vector of observations of length n using maximum likelihood becomes

$$\ln\left(L(\mathbf{X} \mid \mu, \sigma)\right) = \sum_{i=1}^{n} \ln \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2}\right) \tag{18}$$

where, at its maximum value,

$$\hat{\mu} = \bar{x} \qquad\qquad \hat{\sigma} = \sqrt{\frac{(n-1)s^2}{n}} \tag{19}$$

though most researchers will use $\hat{\sigma} = s$ instead as an unbiased estimate.

Under the two sample design with normal population distributions, the maximum likelihood representation becomes

$$\ln\left(L(\mathbf{Y} \mid \mu_C, \sigma_C, \mu_T, \sigma_T)\right) = \sum_{i=1}^{N} \ln\left(\begin{array}{l} h_i \dfrac{1}{\sqrt{2\pi\sigma_C^2}} \exp\left(\dfrac{-(y_i - \mu_C)^2}{2\sigma_C^2}\right) \\[2ex] + (1 - h_i) \dfrac{1}{\sqrt{2\pi\sigma_T^2}} \exp\left(\dfrac{-(y_i - \mu_T)^2}{2\sigma_T^2}\right) \end{array} \right) \tag{20}$$

Because normality or even continuous data are not a requirement, both populations can instead be posited from, say, the Poisson distribution where the likelihood representation becomes

$$\ln\left(L(\mathbf{Y} \mid \lambda_C, \lambda_T)\right) = \sum_{i=1}^{N} \ln\left(h_i \frac{e^{-\lambda_C} \lambda_C^{y_i}}{y_i!} + (1 - h_i) \frac{e^{-\lambda_T} \lambda_T^{y_i}}{y_i!} \right) \tag{21}$$

where **Y** is the concatenated data set and λ_C and λ_T characterize the Poisson distributions for each population. Further, there is no requirement for populations to be the same parametric family. A control group posited from an exponential distribution with the treatment group from the normal distribution is represented as

$$\ln\left(L(\mathbf{Y} \mid \lambda_C, \mu_T, \sigma_T)\right) = \sum_{i=1}^{N} \ln\left(h_i \lambda_C e^{-\lambda_C y_i} + (1 - h_i) \frac{1}{\sqrt{2\pi\sigma_T^2}} \exp\left(\frac{-(y_i - \mu_T)^2}{2\sigma_T^2} \right) \right) \quad (22)$$

Using Equation (20), a series of four models can be constructed, differing only parameter constraints, shown in Figure 2.

Figure 2.
Two Sample Model Representations, Normal Distributions, No Mixing Proportion

TRADITIONAL MODELS

$\phi = 0$

Model #1
$\mu_C \neq \mu_T$
$\sigma_C \neq \sigma_T$
4 Parameters

Model #3
$\mu_C = \mu_T$
$\sigma_C \neq \sigma_T$
3 Parameters

Model #2
$\mu_C \neq \mu_T$
$\sigma_C = \sigma_T$
3 Parameters

Model #4
$\mu_C = \mu_T$
$\sigma_C = \sigma_T$
2 Parameters
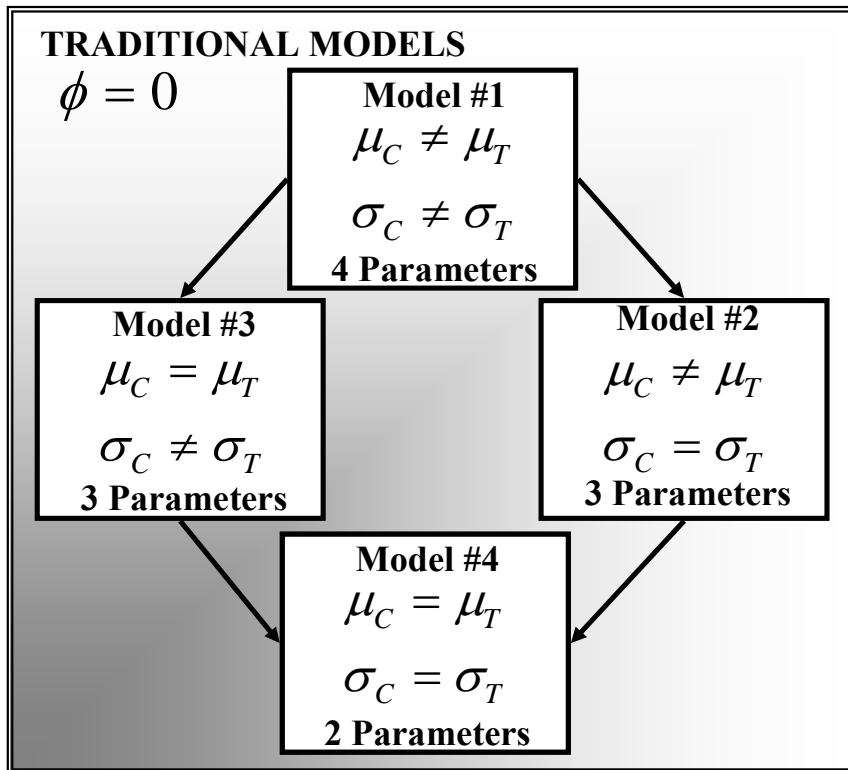
The models represent all possible distribution parameter constraints, freely estimated or constrained across populations, where the figure indicates the total number of estimated parameters for each particular model. The ϕ term, representing a population mixing

15

proportion of treatment nonresponse, is fixed to 0 as represented in Equation (2) as a homogeneous, *ID* sample, while the arrows indicate the hierarchical nesting structure for these models.

If these models were hierarchically nested, a series of likelihood ratio tests and probabilistic inferences could be used to determine the best representation of the data structure. Because Model #2 and Model #3 are not nested, a different comparative method of assessment is required. The lack of nesting among a set of models continues with the introduction of mixture model representations in section 1.3 and subsequently to the more complex extensions in Chapter 4. In a series of influential papers, Hirotugu Akaike advocated an omnibus model assessment measure combining information quantity and a penalty for model parsimony (Akaike, 1973, 1977, 1981). His work maximizes the amount of information from the data and specified model, accomplished by minimizing the Kullback-Leibler function, *-2ln(L)*, with a penalty based on Occam's Razor principle (Kullback, 1951, 1959). Occam's Razor stresses simplicity in model representation, commonly interchanged with the term *parsimony.* There have been a number of challenges to this criterion focused almost exclusively on the size of the parsimony penalty in an attempt to produce a more consistent estimator. As a result, other criterion methods have been proposed including Bayesian/Swartz Information Criteria (*BIC*) (Schwarz, 1978), corrected Akaike Information Criteria (*AICc*), and Hannan-Quinn Information Criteria (*HQC*) (Hannan, 1979). Bozdogan (1987) provided an excellent overview of Akaike's work and connections to these other criteria. The formula for the *AIC* is

$$AIC = 2p - 2\ln(L) \tag{23}$$

16

where ln*(L)* represents the maximum likelihood function value from Equation (20) and *p*

is the number of parameter estimates which varies dependent on model constraints. Of

the four information criterion indices mentioned, only the *AIC* has a constant parsimony

penalty, where as the penalty increases with sample size for the others.

   While simulation studies such as Shibata (1983) and Larimore (1985) have

evaluated these information criteria, there is a lack of theoretical or empirical evidence

generalizing improved model selection for a particular information criterion for the two

sample models presented. For the pilot study presented in section 2.4 and the full study,

AIC is used as the selection criterion. Other criteria could be similarly utilized. One

issue employing information criteria with a sample sized based parsimony penalty in

multi-sample mixture models is use of the total sample size, *N*, as an overly harsh penalty

where in fact different model representations and parameters within may only use subsets

of *N*, in the ML process. This point is discussed in greater detail in section 2.2.

   Maximizing the likelihood function via an appropriate search process such as

Newton-Rhapson, Quasi-Newton, or Sequential Quadratic Programming (SQP), the

likelihood value, *L*, is retained for each model with corresponding parameter estimates.

With datasets containing missing data, another search process called the Expectation-

Maximization (EM) algorithm can be utilized, presented in Dempster, Laird, and Rubin's

(1977) seminal work. After calculating *AIC* values, the lowest *AIC* valued (or min *AIC*)

model is selected as the most viable representation of the population. The number of

parameter estimates and conclusions for each model upon selection are shown in Table 1.

Table 1.

Model Selection Result with Hypotheses' Conclusions, No Mixing Proportion

Model Selected	Parameters Estimated	Hypotheses' Conclusions
1	4	Treatment Effect, Unequal Variances
2	3	Treatment Effect, Equal Variances
3	3	No Treatment Effect, Unequal Variances
4	2	No Treatment Effect, Equal Variances

Model selection results serve as an analog to either version of the t - test in addition to a hypothesis test regarding variance equality. Further, with extensions of multiple treatment means, multiple comparison corrections such as Bonferroni or Scheffé are not required as with probabilistic inference methods. Commonly cited terms as *statistically significant*, Type I error control α, *reject*, and *fail to reject*, are not mentioned with respect to these hypotheses' conclusions. Those terms are artifacts of traditional methods. Probabilistic inferences are not being made; instead, a set of deterministic statements are made based on Akaike Information Criteria values. Qualifying treatment effect and population variance equality results with these terms are not appropriate with this method.

Semantic discussion in usage and appropriateness of such terms obfuscates a valid underlying concern: what is the power of the information criterion method? In other words, how often does this model selection process reach correct hypothesis conclusions? Standard errors for estimates under maximum likelihood determined from the Fisher information matrix might be used as the basis to construct some type of probabilistic inference for individual parameters, but that approach is not easily tractable. This process is in essence a multiple comparison procedure, not a hypothesis test of one model versus another. Further, because all parameters are estimated simultaneously,

considering standard error for a single parameter with different constraints for each

model specification is ill-advised as any type of criterion. Development of an empirical

framework to provide clarity to this question among others presented in section 1.5 is the

sole focus of Chapter 2: Method.

Section 1.3 Nonresponse Impact on Traditional Analysis

An alternative to two statistical tests has been presented using an information-

based model selection process employing a minimum *AIC* strategy. While avoiding

subjective decisions regarding a Type I error control threshold in addition to the variance

equality hypothesis, an information criterion analog is not available in common statistical

software applications. Current strategies addressing treatment nonresponse appear to

reflect this reality. Whether ignoring the possibility of nonresponders or using an ITT

method of post treatment non-probabilistic classification creating distinct groups, both

operate with traditional analysis methods.

Establishing empirical evidence regarding treatment nonresponse presents a

challenge from two aspects. First, a specific assumption of *ID*, identically distributed

observations, is required within the treatment sample for use in fixed effects ANOVA and

either *t* - test procedure. Even the central limit theorem requires an IID sample. Because

one is essentially testing the violation of this assumption, traditional methods cannot be

employed.   Second, common in ANOVA and regression, researchers make inferential

statements regarding model parameters and predictors based on normal distribution

theory as the instrument for empirical evidence. Even if models with some unknown

mixing proportion operated under *ID* conditions, the parameter is bounded [0, 1], therefore not subscribing to a normal distribution.

A comparative evaluation of a series of models, however, is ideally suited to provide empirical evidence. These models assume treatment nonrespondents are distributionally unaffected by application of a particular treatment, retaining the characteristics of the control population. As such, the general representation of the treatment group positing nonresponse becomes

$$X_{Tj} = x_{T1}, x_{T2}, \dots x_{Tn_T} \sim \phi f_C(\boldsymbol{\theta}) + (1 - \phi) f_T(\boldsymbol{\theta}) \tag{24}$$

where ϕ represents the proportion of nonrespondents in the treatment population. To evaluate representations inclusive of mixing proportions requires an additional layer of models, shown in Figure 3.

Figure 3.
Two Sample Model Representations, Normal Distributions, with Mixing Proportion
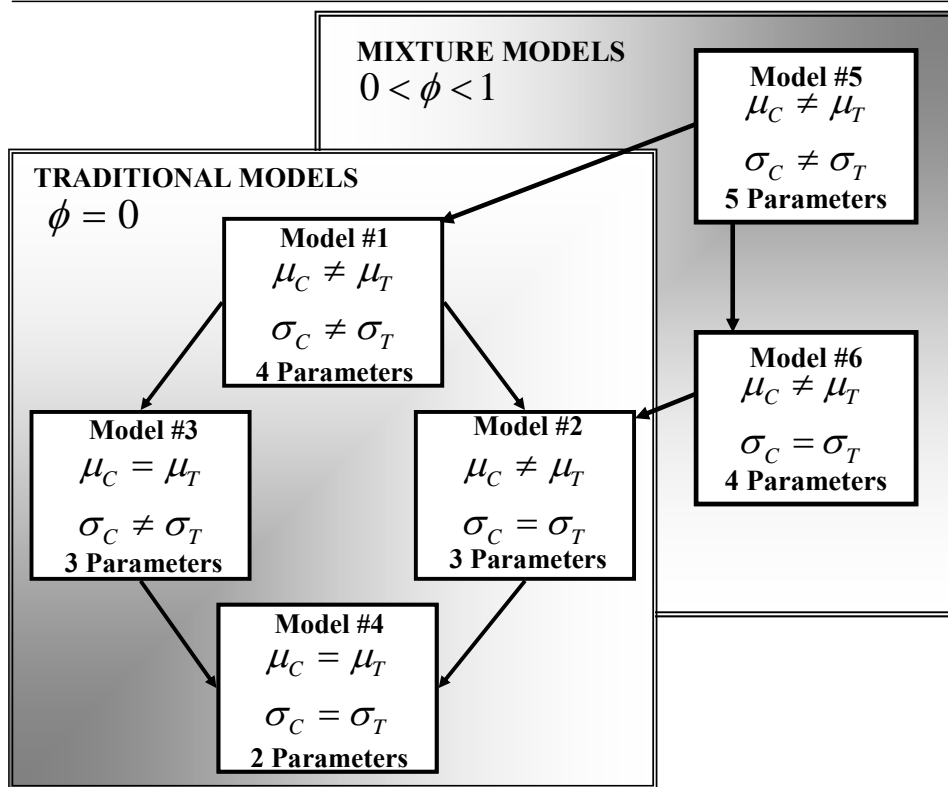
This builds upon Figure 2 introducing two mixture models which require a non-zero

mixing proportion, where arrows indicate the hierarchical nesting structure within and

across these layers. To be thorough, two other model representations are possible with

normal population distributions with a non-zero mixing proportion:

$$\left\langle \begin{array}{l} (7)\ \mu_T = \mu_C, \sigma_T \neq \sigma_C \\ (8)\ \mu_T = \mu_C, \sigma_T = \sigma_C \end{array} \right\rangle \tag{25}$$

Models #7 and #8, while specifying treatment nonresponse, are either not mathematically

tractable or have substantial convergence issues with a ML process. Model #8 results in

a singular matrix with infinitely many solutions at the maximized function value for any

mixing proportion estimate between $[0, 1]$. Successful parameter estimates might be

possible for a Model #7 specification for normal distribution specification given the

independence of their two moments, but there would be a consistency issue. As a more

general statement, convergence of particular representations is much harder for any

parametric distribution whose expected values (means) are constrained to be equal across

the treatment and control populations.

1.4 Mixture Model Framework with Information Criteria

The addition of an extra layer of models shown in Figure 3 supports empirical

evidence for the following hypothesis

$$Ho_1 : \phi = 0 \text{ - Homogeneous (}ID\text{) Population}$$
$$Ha_1 : \phi > 0 \text{ - Treatment Nonresponse Exists}$$

Again, where most hypothesis tests inferentially assess some test statistic in a reject or

fail to reject conclusion, this evidence is information based in a holistic assessment of

model fit. In a number of articles, Dayton similarly advocated a minimum *AIC* strategy

evaluating a series of models (Dayton, 1998, 2003a, 2003b). His motivation, however,

was to replace inferential statements used in multiple sample comparison of population

means, not mixtures, where a dozen post hoc statistic correction procedures exist, at times

with conflicting results. The motivation here is not to replace existing procedures, but

more simply to offer an evidentiary technique addressing treatment nonresponse.

To construct model representations with a mixing proportion, Models #5 and #6

from Figure 3, the treatment sample from Equation (2) is now represented as

$$X_{Tj} = x_{T1}, x_{T2},..., x_{Tn_T} \sim \phi f_C(\boldsymbol{\theta}) + (1-\phi) f_T(\boldsymbol{\theta})$$

$$\text{for } j = 1,2,...,n_T$$

(26)

Inserting this into the maximum likelihood formula, using normal distributions as the

example, results in

$$L(\mu_C, \sigma_C, \phi, \mu_T, \sigma_T) = \prod_{i=1}^{n_C} f_C(x_{Ci} \mid \mu_C, \sigma_C) *$$

$$\prod_{j=1}^{n_T} \left(\phi f_C(x_{Tj} \mid \mu_C, \sigma_C) + (1-\phi) f_T(x_{Tj} \mid \mu_T, \sigma_T) \right)$$

(27)

Simplifying this result using Equations (13) and (14), this formula can be rewritten for

any parametric distribution

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \left[\begin{array}{l} h_i f_C(y_i \mid \boldsymbol{\theta}) + (1-h_i)\phi f_C(y_i \mid \boldsymbol{\theta}) \\ + (1-h_i)(1-\phi) f_T(y_i \mid \boldsymbol{\theta}) \end{array} \right]$$

(28)

where $\boldsymbol{\theta}$ represents the vector of parameter estimates. Maximizing this equation

provides distributional parameters and mixing proportion estimates for a particular model

specification. Because -2ln*(L)* is required for use in the *AIC*, this value can be obtained

simultaneously with corresponding parameter estimates in a single step with the

minimization of

$$-2\ln\left(L(\boldsymbol{\theta})\right) = -2\sum_{i=1}^{N}\ln\left[\begin{array}{l} h_i f_{\text{C}}(y_i \mid \boldsymbol{\theta}) + (1-h_i)\phi\, f_{\text{C}}(y_i \mid \boldsymbol{\theta}) \\ + (1-h_i)(1-\phi)\, f_{\text{T}}(y_i \mid \boldsymbol{\theta}) \end{array}\right] \qquad (29)$$

This general formula is used for all models presented in Figure 3, varying only

distributional parameter and mixing proportion constraints. This formula also applies to

any parametric distribution specifications; however, the number and construction of

possible models will differ as subsequently illustrated in Chapter 4. Selecting the min

AIC from among the set of models enables a treatment nonresponse hypothesis

conclusion building on the results from Table 1 shown in Table 2.

Table 2.
Model Selection Result with Hypotheses' Conclusions With Mixing Proportion

Model Selected	Parameters Estimated	Hypotheses' Conclusions
1	4	Homogeneous Population, Treatment Effect, Unequal Variances
2	3	Homogeneous Population, Treatment Effect, Equal Variances
3	3	Homogeneous Population, No Treatment Effect, Unequal Variances
4	2	Homogeneous Population, No Treatment Effect, Equal Variances
5	5	Treatment Nonresponse, Treatment Effect, Unequal Variances
6	4	Treatment Nonresponse, Treatment Effect, Equal Variances

Certain research fields and journals might require inferential test statistic values

and *p*-values for parameter estimates, but that is not the mean of evidentiary support

utilized with this approach. Distributional parameter and mixing proportion estimates

should be reported, though should a researcher feel compelled to provide such a statistic,

the following may be used

$$\hat{z}^* = \frac{\mid \hat{\mu}_{\text{C}} - \hat{\mu}_{\text{T}} \mid}{\sqrt{\dfrac{\hat{\sigma}_{\text{C}}^2}{n_{\text{C}} + \hat{\phi}\, n_{\text{T}}} + \dfrac{\hat{\sigma}_{\text{T}}^2}{(1-\hat{\phi})n_{\text{T}}}}} \qquad (30)$$

associated with inferential statements regarding population mean differences. The

corresponding *p*-value is

$$p - \text{value} = \Phi^{-1}(\hat{z}^*) \qquad (31)$$

While Equation (30) uses normal distributions, this statistic can be used for any specified population parametric distributions capitalizing on the central limit theorem, given sufficient sample sizes, with the general form

$$\hat{z}* = \frac{| E(\mathbf{X} | \hat{\boldsymbol{\theta}}_\text{C}) - E(\mathbf{X} | \hat{\boldsymbol{\theta}}_\text{T}) |}{\sqrt{\dfrac{V(\mathbf{X} | \hat{\boldsymbol{\theta}}_\text{C})}{n_\text{C} + \hat{\phi} n_\text{T}} + \dfrac{V(\mathbf{X} | \hat{\boldsymbol{\theta}}_\text{T})}{(1 - \hat{\phi}) n_\text{T}}}} \tag{32}$$

Previously noted, the impact of treatment nonresponse not only confounds estimation of the treatment population mean, but disqualifies traditional analysis statistical testing because of an *ID* violation in the treatment sample. Both Equations (30) and (32) indicate probabilistic inferences remain problematic when considering sample sizes. Sample size, often ignored as a readily available input for statistical tests, is replaced by *effective* sample size which can only be estimated. The use of the term effective sample size is different in this application compared with Louis' use of the same term associated with the EM algorithm (1982). Used in the post model selection test statistic, replacing a known n_C and n_T used in section 1.1, these effective sample sizes are

$$n_\text{C}^* = n_\text{C} + \hat{\phi} n_\text{T} \tag{33}$$

and

$$n_\text{T}^* = (1 - \hat{\phi}) n_\text{T} \tag{34}$$

where the total sample size, *N*, remains unchanged

$$N = n_\text{C}^* + n_\text{T}^* = n_\text{C} + \hat{\phi} n_\text{T} + (1 - \hat{\phi}) n_\text{T} \tag{35}$$

A final comment regarding two sample designs involving repeated measures using this method is noteworthy. With traditional analyses, an assumption of normality was required for baseline and post treatment measurements. The interest in this approach

24

is neither the baseline nor post treatment measurement, but their difference as some parametric form, normally distributed or otherwise. Differencing creates the independence necessary by combining measurements for an individual. The result is population models with and without mixtures using the differenced measure of an individual, where the distributional structure of its initial components is irrelevant. No probabilistic inferences are made, normality is not required, and any parametric distribution(s) may be posited for these populations and used with Equation (29).

1.5 Empirical Research Questions

The purpose of the proposed empirical research, like most others, is to evaluate the performance of a particular statistic, correction, or in this case methodology under controlled conditions. Despite all the formulas and models presented, the systematic process is straightforward, outlined in the following steps:

a. For a set of population conditions, generate samples of data where:

   1) The mixing proportion, ϕ, assumes zero and non-zero values.

   2) Nonrespondents in the treatment group have the same distributional properties as the control population.

b. For each sample of data, fit the six models within a ML framework and using a min *AIC* strategy, select the most representative model.

c. For each selected model, retain a series of performance metrics.

d. Steps (a) through (c) constitute a single trial. Repeat for many trials, retaining summary information for all trials for the particular set of conditions.

e. Repeat steps (a) through (d) for all other population conditions.

Because there is no precedent for an empirical study of this nature, most elements within this systematic process require considerable development. For instance, it is unreasonable to assume conditions varied in traditional studies of test statistic performance directly map to information based model selection process involving mixtures. The phrase "series of performance metrics" from (c) above is a bit vague, where unlike evaluating a particular statistic, this evaluation is on an entire methodological process. Finally, because an optimization algorithm is utilized in parameter estimation, certain technical elements can affect the performance. Accordingly, Chapter 2 Method will be a bit unconventional in its development.

The empirical research questions explored in this dissertation target three different levels: model, parameter, and individual. At the model level,

    a. How often does this process select the correct model?

    b. How often does this process provide correct empirical evidence regarding differences in population means?

    c. How often does this process provide correct empirical evidence regarding treatment nonresponse?

    d. How often does this process provide correct empirical evidence for equality of variances?

    e. For models which require an optimization algorithm, what is the percentage of viable convergence per model per experimental condition?

Parameter level questions focus on the min *AIC* selected model, irrespective of whether it was the correct population representation, include:

a. What is the *bias* and mean squared error (*MSE*) for the population mixing

   proportion estimate?

b. What is the *bias* and *MSE* for the control and treatment population

   distributional parameter estimates?

Research questions at the individual level, focused on respondents from the treatment

sample for a chosen model, are:

a. What is the individual average error in probability of class membership as a

   treatment nonresponder?

b. What is the overall percentage of correct classification for individuals within a

   treatment sample?

Chapter 2: Method

This chapter seeks clarity to research questions presented in section 1.5 under controlled experimental conditions, focusing on two sample designs using normal distributions. Even though the methodology supports other parametric distribution specifications, normality in both populations remains the example in technical discussions, mathematic representations of performance criteria, development of control variables, and use in the pilot study (the progression of this chapter).

Unfortunately, no currently available software performs the model comparison procedure presented in Chapter 1, so code was developed in Gauss 8.0 (Aptech, 2005). Code results were validated, where possible, with EQS 6.1 (Bentler, 2006), Stata 10 (StataCorp, 2007), and using Microsoft Excel premium solver. More specific information with regard to code validation is provided in the next section with other technical information requisite for code construction.

An empirical study to evaluate these research questions not only entails identification of influential study parameters for systematic variation, but their upper and lower boundaries and segmentation within where substantive changes in performance occur. Without similar studies in the literature for reference, a degree of theoretical development is required. An ensuing pilot study not only validates study parameter selection, but provides insight regarding parameter boundaries prior to engaging in a comprehensive study. The Pavlic study is the closest in conception, having conducted simulations under a two sample design with treatment nonresponse. Pavlic simulated fit of a Gaussian mixture under 90 different conditions of 100 trials each, only fitting the

correct population model. Experimental conditions evaluated had very large, equal sample sizes, 1000 and 2000 in each sample, with unequal population variances in a fixed 1.25 ratio. These qualifications are offered not to diminish the quality of his study, but to provide some inkling to the complexity and computational intensiveness of a comprehensive multi-model selection study. By comparison, for each sample of data, six models are fit instead of only the correct specification. Further, approximately 580 experimental conditions with 500 trials per condition are evaluated, a six and five fold respective increase.

2.1 Technical Discussion

Software enabling a min *AIC* selection of competing models involving mixtures is unavailable for even a single set of data, let alone to support an extensive empirical study. In supporting this effort, code was developed in Gauss 8.0, which can be made available upon request. To validate parameters estimates, $-2\ln(L)$, and *AIC* values of the code output, two fixed sample datasets were compared against Microsoft Excel premium solver for all six models. Results from Models #1-4 were also validated against EQS 6.1, which does not accommodate mixture representations. For Models #5-6, considering *only* the treatment group sample, the *denormix* (Kolenikov, 2001) add-on package for Stata verified the results. A number of other stand alone software choices, including MPlus (Muthen, 2001) and others evaluated by Haughton (1997), which allow specification of other parametric distributions, are available, but focus on extracting mixtures from single data samples, not multi-sample designs. Liesch (2004) provides a flexible add-on package for R (R, 2008) called *flexmix* supporting regression and a

broader family of general linear models, though without support to multi-sample designs. MPlus does allow constraints across groups in mixture model representations, but still does not support this methodological process in the univariate case or the multivariate latent extension discussed in Chapter 4 where its limitations are detailed.

A number of technical decisions were required in the construction of code to support the empirical study.

a. Data were generated according to distribution population parameters, drawing samples of size n_C and n_T, with a specified population mixing proportion, ϕ, present within the treatment sample. Fixing this value in the sample instead of the population from which a sample was drawn made ϕ consistent for each trial. For example, experimental conditions with ϕ of 0.20 in a n_T of 100 always has 20 treatment nonresponders in each treatment sample whose response subscribed to the control population distribution.

b. 500 trials were conducted for each experimental condition. A sufficient number of trials are required given the interest in percentages of model selection supporting each hypothesis. The choice of 500 trials is tempered against the extensive computational requirement fitting six models, five of which are misspecified, requiring a non-linear optimization algorithm in three models. Pilot study results, depending on experimental conditions, required between 2 hours and 3 days to complete 500 trials on a dedicated Pentium P4 3.0 GHz with 2GB RAM running Windows XP with the Gauss system cache increased from 32KB to 512KB.

c. The non-linear optimization command within Gauss, *SQPSolve*, only supports determination of a minimum value. This conveniently coincides with Equation (29) enabling direct calculations of the -2ln*(L)* value.

d. The *SQPSolve* command allows parameter bounds to be specified to quicken the optimization search process. Population standard deviation estimates were restricted to be greater than zero, avoiding a degeneracy issue noted by Hathaway (1985) and Ridolfi (1999). For mixture models, the population mixing proportion is restricted to be greater than 0 and less than 1. Optimization estimate results of 0 and 1 are redundant, corresponding to traditional Models #1 and #4 respectively.

e. Initially, convergence of the mixing proportion estimate was problematic using the *SQP* process given its natural restriction of values between 0 and 1. Despite having the same likelihood shape, rescaling the mixing proportion parameter within the likelihood function ranging between 0 and 1000 alleviated this problem . For example, an output of 426 corresponds to a mixing proportion estimate of 0.426.

2.1.1 Model Estimation Techniques

Three of the models utilize mathematically proven results foregoing the need for any optimization algorithm. Incorporating these results directly into the code created a more efficient application that reduced processing time. Model #1 posits no mixing proportion, requires four parameters estimates, where the minimum value of the -2ln*(L)* function occurs at

$$\hat{\mu}_{\mathrm{C}} = \bar{x}_{\mathrm{C}} \qquad\qquad \hat{\mu}_{\mathrm{T}} = \bar{x}_{\mathrm{T}} \qquad\qquad (36)$$

$$\hat{\sigma}_{\mathrm{C}} = \sqrt{\frac{(n_{\mathrm{C}} - 1)s_{\mathrm{C}}^2}{n_{\mathrm{C}}}} \qquad \hat{\sigma}_{\mathrm{T}} = \sqrt{\frac{(n_{\mathrm{T}} - 1)s_{\mathrm{T}}^2}{n_{\mathrm{T}}}} \qquad\qquad (37)$$

Model #2 similarly posits no mixing proportion, but requires three parameter estimates where the minimum value of the -2ln(L) function occurs with Equation (36) and a single standard deviation estimate of

$$\hat{\sigma} = \hat{\sigma}_{\mathrm{C}} = \hat{\sigma}_{\mathrm{T}} = \sqrt{\frac{n_{\mathrm{C}} s_{\mathrm{C}}^2 + n_{\mathrm{T}} s_{\mathrm{T}}^2}{N}} \qquad\qquad (38)$$

which is weighted averaged of the sample variances. Model #3 also has three parameters, but requires a non-linear optimization of Equation (29) with the following constraints

$$\phi = 0, \ \mu = \mu_{\mathrm{C}} = \mu_{\mathrm{T}} \qquad\qquad (39)$$

where estimates

$$-2\ln(L), \ \{\hat{\mu}, \hat{\sigma}_{\mathrm{C}}, \hat{\sigma}_{\mathrm{T}}\} \qquad\qquad (40)$$

are returned. Model #4 has two parameters where the minimum value of the -2ln(L) occurs at

$$\hat{\mu} = \hat{\mu}_{\mathrm{C}} = \hat{\mu}_{\mathrm{T}} = \bar{y} \qquad\qquad (41)$$

$$\hat{\sigma} = \hat{\sigma}_{\mathrm{C}} = \hat{\sigma}_{\mathrm{T}} = \sqrt{\frac{(N-1)s_y^2}{N}} \qquad\qquad (42)$$

where **Y** represents both samples in a concatenated data set presented in Equation (13). Model #5 has five parameters, requiring an optimization of equation (29) with the following constraint

$$0 < \phi < 1 \qquad\qquad (43)$$

where estimates

$$-2\ln(L), \left\{\hat{\phi}, \hat{\mu}_\mathrm{C}, \hat{\sigma}_\mathrm{C}, \hat{\mu}_\mathrm{T}, \hat{\sigma}_\mathrm{T}\right\} \qquad (44)$$

are returned. Model #6 also includes a mixing proportion, though only has four

parameters where the constraints in the optimization of Equation (29) are

$$0 < \phi < 1, \sigma = \sigma_\mathrm{C} = \sigma_\mathrm{T} \qquad (45)$$

returning estimates

$$-2\ln(L), \left\{\hat{\phi}, \hat{\mu}_\mathrm{C}, \hat{\mu}_\mathrm{T}, \hat{\sigma}\right\} \qquad (46)$$

SQPSolve employs a sequential quadratic programming (SQP) search method, a

generalization of the more common Newton's method. Default settings for the tolerance

condition of 1E-5 and the maximum number of search iterations of 1E+5 were retained.

The success of any optimization routine depends, in part, on initial parameter estimates,

commonly called starting values, initiating the search process. These values not only

affect convergence, but arrival at local optimal solutions opposed to a true global

solution.

     For the three models requiring an optimization algorithm for parameter

estimation, the same process was used to generate starting values regardless of population

conditions. Using Model #3 for example, Equation (41) is used as the representative

measure for a single population mean. Similarly, for Model #6, Equation (38) is used as

the representative measure for the same standard deviation across populations. For any

particular set of starting values, the algorithm may not converge as defined in the next

section. In such cases, new starting values were created and the process was repeated.

Let $\hat{\boldsymbol{\theta}}_\mathrm{P}$ represent a vector of summary statistics commensurate with a particular model

specification for its distributional parameters, corresponding starting values are

determined by

$$\hat{\mathbf{\theta}}_{\text{Initial}} = (1 + .075 * \text{Rand}_{\Phi})\hat{\mathbf{\theta}}_{\text{P}} \qquad (47)$$

There are no summary measures available to serve as a reasonable starting point for the population mixing proportion, so

$$\hat{\phi}_{\text{Initial}} = \text{Rand}_{\text{Uni}}(0.01, 0.65) \qquad (48)$$

was used. Treatment population summary measures become increasingly inaccurate with higher proportions of nonresponse, though still serve as a reasonable starting point upon which to initiate the algorithmic search process. The random generation process embedded in the creation of starting values is necessary to mitigate convergence issues.

2.1.2 Convergence

The term *convergence* in this dissertation encompasses the conventional definition of the search algorithm successfully arriving at a set of estimates with a more restrictive second condition of satisfying specific model requirements. Whether a result of poor starting values, model complexity, or model misspecification, an optimization algorithm can fail to converge in a computational manner for such reasons as gradient, Hessian, and function calculation failures, or exceeding the maximum number of iterations. The second condition of satisfying model specific requirements capitalizes upon the hierarchical nesting structure shown in Figure 3.

Regardless of the population conditions underlying the sample of data, Models #3, #5, and #6 require use of an optimization algorithm. For each model, if the process does not reach a solution or the solution fails to meet model fit conditions defined subsequently, a failed attempt is registered. Each model is afforded up to 15 attempts to attain a viable set of parameters estimates. While such problems can also occur for an

applied researcher, it is greatly exacerbated in an empirical study which contributed to the often lengthy amount of time necessary to complete 500 trials. Models failing to converge after 15 attempts are assigned an excessively high AIC value essentially eliminating their selection in a min *AIC* comparison. Because the process ceases when satisfying model fit conditions, the -2ln*(L)* value is assumed to be the global minimum. More rigorous methods ensuring a global minimum are presented in Chapter 4, though are better applied with a single data set. Employing these methods requires additional computations, and given the significant time required for a set of trials, was not implemented in this study.

Models #3, #5, and #6 must also satisfy model fit conditions taking advantage of their hierarchical relationships with models that do not require an optimization algorithm for estimation. Model #3 results must meet the following condition

$$-2\ln(L)_{\text{Model1}} \leq -2\ln(L)_{\text{Model3}} \leq -2\ln(L)_{\text{Model4}} \tag{49}$$

as Model #3 is hierarchically nested between Model #1 and Model #4. Model #1 and model #4 use established results to obtain to their -2ln*(L)* values. For Model #5

$$-2\ln(L)_{\text{Model5}} \leq -2\ln(L)_{\text{Model1}} \tag{50}$$

while for Model #6

$$-2\ln(L)_{\text{Model6}} \leq -2\ln(L)_{\text{Model2}} \tag{51}$$

must be satisfied as a result of their hierarchical relationships with Models #1 and #2 respectively. The hierarchical relationship between Models' #5 and #6 is not used because both require an optimization algorithm and a failed convergence and resulting high *AIC* assignment would be problematic when evaluating model fit conditions.

2.2 Simulation Parameter Development

Without the ability to reference similar studies in the literature, identification of controlled study parameters begins with some theoretical development. Focusing on the central element of the *AIC*, the -2ln*(L)* value, rewriting Equation (29)

$$-2\ln\big(L(\boldsymbol{\theta})\big) = -2\sum_{i=1}^{N} \ln\left[\begin{array}{l} h_i f_C(y_i \mid \boldsymbol{\theta}) + (1-h_i)\phi\, f_C(y_i \mid \boldsymbol{\theta}) \\ + (1-h_i)(1-\phi)\, f_T(y_i \mid \boldsymbol{\theta}) \end{array}\right]$$

becomes the following with normal population distribution specifications

$$-2\ln\big(L(\mu_C,\sigma_C,\phi,\mu_T,\sigma_T)\big) = -2\sum_{i=1}^{n_C+n_T} \ln\left[\begin{array}{l} h_i f_C(y_i \mid \mu_C,\sigma_C) \\ + (1-h_i)\phi\, f_C(y_i \mid \mu_C,\sigma_C) \\ + (1-h_i)(1-\phi)\, f_T(y_i \mid \mu_T,\sigma_T) \end{array}\right] \qquad (52)$$

Simple inspection reveals seven variables (five population parameters and two sample size conditions) that require variation

$$\left(n_C,\mu_C,\sigma_C,n_T,\mu_T,\sigma_T,\phi\right) \qquad (53)$$

Even the coarsest manipulation of these variables, two levels each, totals 128 experimental conditions. Such coarse levels of manipulation provide a very limited ability to generalize findings and characterize relationships in order to assess the posited research questions. Consideration of more levels per variable becomes a combinatoric expansion unsupportable given the time length to complete 500 trials. This eliminates a multifactorial simulation approach without some parameter reduction. The goal in development, therefore, is reduction of a seven parameter space to a more manageable three parameter set by combining variables in a theoretically supported manner. This reduction comparatively enables greater variation of controlled study parameters resulting in a more informative set of experimental conditions. Identification of variable

36

composites also affects population level research questions, replacing individual parameter evaluations as discussed in the next section. Brief assessments for each of these seven variables, interrelationships, and impact on Equation (52), are provided below:

    a. Sample sizes, n_C and n_T: The control group sample size, n_C, affects the accuracy of the control population parameter estimates $\{\hat{\mu}_C, \hat{\sigma}_C\}$. The treatment group sample size, n_T, affects the accuracy of the treatment population parameter estimates $\{\hat{\mu}_T, \hat{\sigma}_T\}$. Both are conditioned upon correct model specification.

    b. Mixing Proportion, ϕ: For models specifications including a mixing proportion, no longer are n_C and n_T independently responsible for determining respective population distributional parameters, instead replaced by *effective* sample sizes shown in Equations (34) and (35). Introduced in section 1.4 and further detailed in the next section, these terms require estimation. Because population conditions are controlled in this study, the *true* effective sample sizes, n_C^* and n_T^* are known. Accuracy of the mixing proportion estimate is influenced by the total sample size, N. Its accuracy is also influenced by a treatment effect; a measure, subsequently presented, summarizing a degree of separation between control and treatment populations.

    c. Control population parameters, $\{\mu_C, \sigma_C\}$: Estimates for these parameters and their accuracy are determined by some combination of sample or effective sample sizes depending on model specification. For example, the Model #1

$\hat{\mu}_C$ estimate is influenced only by n_C where for Model #5, $\hat{\mu}_C$ is influenced

by n_C^* (or $n_C + \phi n_T$).

    d. Treatment population parameters, $\{\mu_T, \sigma_T\}$: The same points discussed in the

       control population are applicable whose estimate accuracy is contingent on

       effective sample size and model specification. For example, the Model #2 $\hat{\mu}_T$

       estimate is influenced by n_T.

Resulting from these assessments, the following parameters were controlled in an

empirical study, further explicated in subsequent subsections

$$\{\text{treatment effect}, \phi, n_T\} \tag{54}$$

2.2.1 Treatment Effect

Defined as a measure of separation between the control and treatment

populations, a treatment effect impacts the accuracy of the mixing proportion estimate.

Of the three presented, the simplest is an unstandardized treatment effect defined as

$$d_u = |\mu_C - \mu_T| \tag{55}$$

This representation of treatment effect is unit dependent and becomes difficult to

generalize across experiments where mean representations operate on different scales. A

more common representation is a standardized treatment effect such as Cohen's d (1988),

represented as

$$d = \frac{|\mu_C - \mu_T|}{\sqrt{\dfrac{n_C \sigma_C^2 + n_T \sigma_T^2}{N}}} \tag{56}$$

where, after adjusting for effective sample sizes, takes the form

$$d^* = \frac{|\mu_C - \mu_T|}{\sqrt{\dfrac{n_C^* \sigma_C^2 + n_T^* \sigma_T^2}{N}}} = \frac{|\mu_C - \mu_T|}{\sqrt{\dfrac{(n_C + \phi n_T)\sigma_C^2 + (1-\phi)n_T \sigma_T^2}{N}}} \qquad (57)$$

This unitless measure generalizes well across mean and variance combinations and is not influenced by particular n_C and n_T values, only their ratio. Further, this standardized treatment effect is often reported in scholarly and applied publications perhaps due to its easy interpretation. Aside from differences in population means, variation in the other parameters in a two sample design can also occur in the following configurations:

Type 1: $\sigma_C \neq \sigma_T, n_C \neq n_T$
Type 2: $\sigma_C \neq \sigma_T, n_C = n_T$
Type 3: $\sigma_C = \sigma_T, n_C \neq n_T$
Type 4: $\sigma_C = \sigma_T, n_C = n_T$

Exploring this treatment effect measure in relation to these elements and the mixing proportion, let

$$a = n_C / n_T \qquad (58)$$

where a represents the sample size ratio and

$$b = \sigma_{rat}^2 = \sigma_C^2 / \sigma_T^2 \qquad (59)$$

where b represents the variance ratio. Substituting these results into Equation (57) produces

$$d^* = \frac{|\mu_C - \mu_T|}{\sqrt{\dfrac{(an_T + \phi n_T)b\sigma_T^2 + (1-\phi)n_T \sigma_T^2}{an_T + n_T}}} \qquad (60)$$

which after collecting like terms and some reduction, yields

$$d^* = \frac{|\mu_C - \mu_T|}{\sqrt{\left(\dfrac{ab + \phi b + 1 - \phi}{a+1}\right)\sigma_T^2}} \qquad (61)$$

reaffirming this measure is not a function of either sample size. Further, when the

population variances are equal ($b = 1$), this reduces further to

$$d^* = \frac{|\mu_{\mathrm{C}} - \mu_{\mathrm{T}}|}{\sigma} \tag{62}$$

where this measure is now unaffected by the mixing proportion.

There is a third representation of treatment effect, possibly not thought of in such

terms. Using normal distributions specifications, the difference of two population means

expressed as z^*, inclusive of a mixing proportion, is

$$z^* = \frac{|\mu_{\mathrm{C}} - \mu_{\mathrm{T}}|}{\sqrt{\dfrac{\sigma_{\mathrm{C}}^2}{n_{\mathrm{C}} + \phi n_{\mathrm{T}}} + \dfrac{\sigma_{\mathrm{T}}^2}{(1 - \phi)n_{\mathrm{T}}}}} \tag{63}$$

This is also a unitless measure, though unlike d^*, has a dependence on sample size. This

measure can be rewritten to accommodate any parametric distribution where

$$z^* = \frac{|\mathrm{E}(\mathbf{X}_{\mathrm{C}}) - \mathrm{E}(\mathbf{X}_{\mathrm{T}})|}{\sqrt{\dfrac{\mathrm{V}(\mathbf{X}_{\mathrm{C}})}{n_{\mathrm{C}} + \phi n_{\mathrm{T}}} + \dfrac{\mathrm{V}(\mathbf{X}_{\mathrm{T}})}{(1 - \phi)n_{\mathrm{T}}}}} \tag{64}$$

Despite similarities in construction with the Satterwaite test statistic, it is not used for

probabilistic inferences on estimated parameter differences.

Unfortunately, while controlling for a treatment effect in the study, Equation (52)

cannot be formally expressed as a function of d_u, d^*, or z^*, requiring empirical validation

accomplished via a pilot study. To illustrate differences and select from among these

treatment effect representations, consider the following table indicating whether a

treatment effect measure is affected by a particular change in experimental conditions.

Table 3.
Comparison of Treatment Effect Measures

Variance Equality	Condition Change	Treatment Effect Affected (Y/N)?		
		d_U	$d*$	$z*$
Any condition	n_C or n_T sizes	NO	NO	YES
$\sigma_C^2 = \sigma_T^2$	μ_C or μ_T	YES	YES	YES
	ϕ	NO	NO	YES
	Sample size ratio	NO	NO	YES
$\sigma_C^2 \neq \sigma_T^2$	μ_C or μ_T	YES	YES	YES
	ϕ	NO	YES	YES
	Sample size ratio	NO	YES	YES
	Variance ratio	NO	YES	YES

As Table 3 illustrates, no treatment effect measure is immune to every experimental condition change. Because of its consistency in changing across all conditions, $z*$ will be used as the controlled parameter representative of treatment effect. This does not mean the other treatment effect measures are neglected; to the contrary, shown in the next section, population level research questions focus on recapturing both d_u and $d*$.

Figures 4 and 5 graphically illustrate differences in standardized treatment effect measures when the population variances are equal. For both figures, the d^* measure remains constant at 2 showing the comparative z^* value when total sample size, sample size ratio, and mixing proportion conditions are varied.

Figure 4.
Graph: Treatment Effect Relationship with Total Sample Size and Sample Ratio
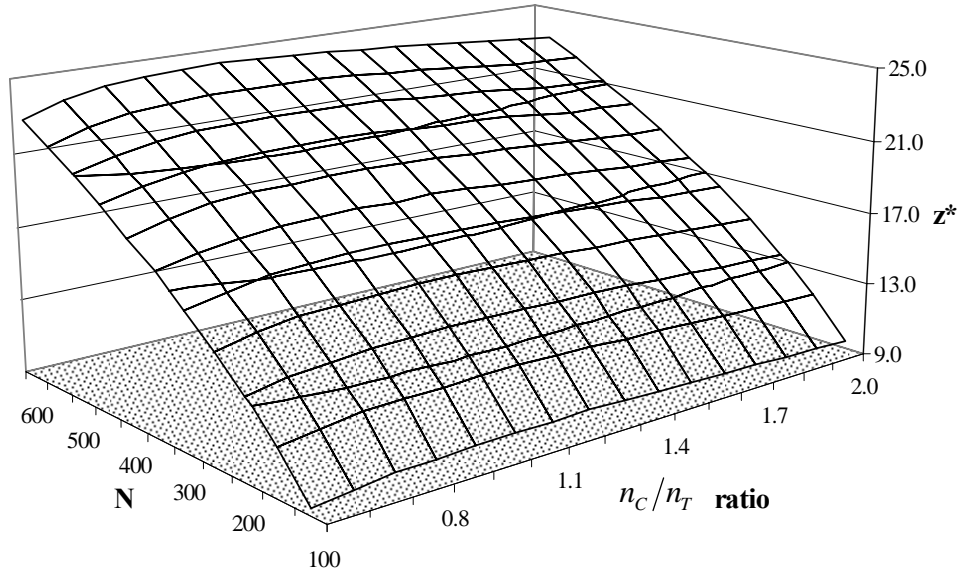
Figure 5.
Graph: Treatment Effect Relationship with ϕ and Sample Size Ratio
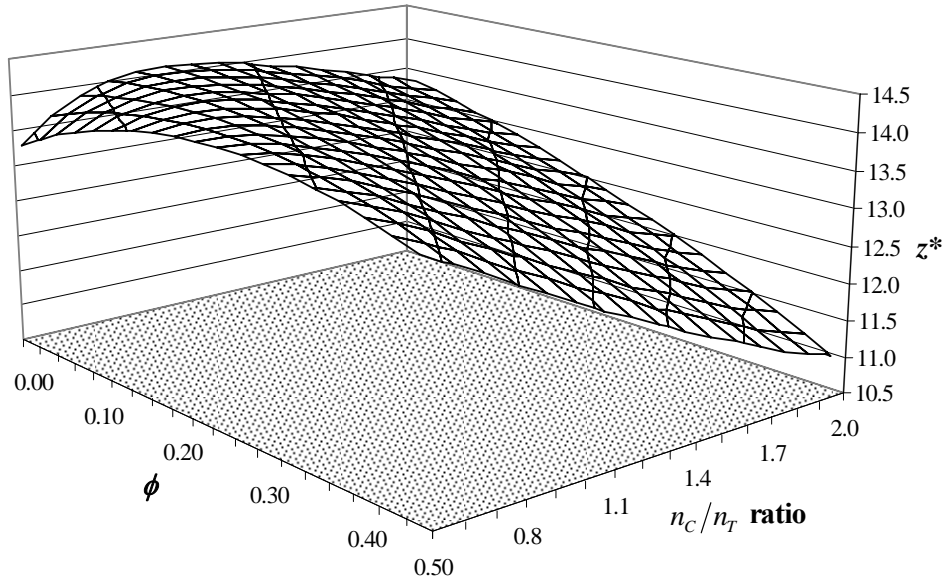
If suspicious of treatment nonresponse, d and d_u estimates should not be

calculated by using sample summary measures. Use of these summary measures as

estimators for the treatment population mean and variance produce inaccurate, negatively

biased results and invalidate the statistical tests presented in section 1.1. To quantify

these inaccuracies, if a researcher fails to consider treatment nonresponse, treatment

effect estimates from the summary measures, referred to as $\mu_{\text{T\_ID}}$ and $\sigma_{\text{T\_ID}}$ because of

the incorrectly assumed ID property become

$$\mu_{\text{T\_ID}} = \text{E}(\mathbf{X}_{\text{T}}) = \phi\mu_{\text{C}} + (1-\phi)\mu_{\text{T}} \tag{65}$$

and

$$\sigma_{\text{T\_ID}} = \sqrt{\phi\sigma_{\text{C}}^2 + (1-\phi)\sigma_{\text{T}}^2 + \phi(1-\phi)(\mu_{\text{C}} - \mu_{\text{T}})^2} \tag{66}$$

The impact of utilizing the summary measures is two fold: it underestimates the *true*

treatment population mean and overestimates the true treatment population variance.

This results in expected negative biases for the treatment effect estimate. With equal

sample sizes, the unstandardized treatment effect *bias* is

$$\text{E}(bias(d_u)) = -\phi|\mu_{\text{C}} - \mu_{\text{T}}| \tag{67}$$

where the assumed IID summary statistics underestimate the unstandardized treatment

effect by up to 50% in the experimental conditions evaluated.

The degree of negative bias for the standardized treatment effect, Cohen's *d*, is

more severe given the overestimation of the treatment population variance. Using equal

sample sizes, the *bias* is represented as

$$\text{E}(bias(d)) = \frac{|(1-\phi)(\mu_{\text{C}} - \mu_{\text{T}})|}{\sqrt{\dfrac{\sigma_{\text{C}}^2 + \phi\sigma_{\text{C}}^2 + (1-\phi)\sigma_{\text{T}}^2 + \phi(1-\phi)(\mu_{\text{C}} - \mu_{\text{T}})^2}{2}}} \\ - \frac{|\mu_{\text{C}} - \mu_{\text{T}}|}{\sqrt{\dfrac{(1+\phi)\sigma_{\text{C}}^2 + (1-\phi)\sigma_{\text{T}}^2}{2}}} \tag{68}$$

underestimating the standardized treatment effect by up to 70% in the study's

experimental conditions. For example, the consequences of using summary measures

are quite harsh where a true standardized treatment effect of 1.0 instead would

be estimated at 0.30.

2.2.2 Mixing Proportion:

Of all parameters requiring estimation, this is arguably the most important. If

group membership were truly known to a researcher, conventional statistical tests for

comparing population mean differences become available and probabilistic inferences on

the mixing proportion estimate are not necessary. Because that is not the case, we are

relying on an omnibus information-based approach as the form of empirical evidence for

both research interests.

Choice of the mixing proportion as a controlled study parameter implies

performance of this methodology is additionally influenced beyond integration into the

treatment effect measure calculation. While accuracy of the mixing proportion estimate

is a function of treatment effect and sample size, this process is not wholly about

accuracy in parameter estimation; it is first and foremost an issue of model selection to

obtain empirical evidence. Of course, an inherent belief is that greater accuracy of

parameter estimates in an omnibus sense results in a greater likelihood in of identifying

the correct model specification. From a practical matter in analyzing the *AIC*, Equation

(23), the improvement of model fit from an additional parameter must exceed the

parsimony penalty. For example, for model specifications differing only by inclusion of

a mixing proportion, the resulting *AIC* improvement must be greater than 2 to select the

more complex representation. It is possible, therefore, for population values to be exactly

estimated under a correct specification and that model is not selected. As the true mixing

proportion approaches 0, holding sample size and treatment effect constant, there is

greater likelihood the overall model fit improvement will not exceed 2.

This is an appropriate time to readdress selection of the *AIC* from among other

possible information criteria. The *AIC* is the only criterion whose parsimony penalty is

independent of sample size, and with sample sizes considered in the study it is the most

favorable for selection of more complex mixture models. While the literature is silent

with regard to an information criterion preference in multi-sample mixture models, there

is an issue in implementing sample size based parsimony penalties. Demonstrating this

issue, Table 4 shows each model and the corresponding sample sizes used to estimate

their particular parameters.

Table 4.
Effective Sample Sizes for Parameters in Specified Models

Model	Parameter				
	ϕ	μ_C	σ_C	μ_T	σ_T
1	None	n_C	n_C	n_T	n_T
2	None	n_C	N	n_T	None
3	None	N	n_C	None	n_T
4	None	N	N	None	None
5	N	n_C^*	n_C^*	n_T^*	n_T^*
6	N	n_C^*	N	n_T^*	None

The use of total sample size is not precise as parameters within a model use smaller and

different effective sample sizes in their estimation. Nor is there a single representative

sample size for all parameters for any model enabling construction of a consistent

penalty. Using effective sample sizes determining the parsimony penalty would impose

different penalties for the same data depending on model specification. Also unclear with

a sample sized based parsimony penalty is resolution when the EM algorithm is required

due to incomplete datasets. Discussion of *AIC* and other information criteria is again

revisited in the context of the study results, section 3.4.

2.2.3 Treatment Group Sample Size

Model parameter estimates are determined from the information resident the data

samples, albeit in different configurations based on the model specification illustrated in

Table 4. The greater the effective sample size, the less the impact of sampling error.

While z^* requires sample sizes for its calculation, it does not fix these values. Selection

of a treatment group sample size, n_T, as a controlled study parameter fixes values for an

empirical study where determination of total sample size, N, and control group sample

size, n_C, are attained using the n_C / n_T ratio defined in Equation (58). Another aspect of

sample size is its effect on the fixed parsimony penalty in the *AIC*.   Larger total sample

sizes invariably result in higher -2ln*(L)* values where a fixed penalty has comparatively

smaller impact to an overall fit measure in competing models.

2.3 Performance Measures

With study parameters chosen for systematic manipulation, attention turns to

providing mathematical specificity to accompany the research questions. Maintaining the

same structure as the research questions in section 1.5, corresponding performance

measures are introduced in a top down approach: model, population, and individual level.

2.3.1 Model Level:

For any experimental condition, retaining the proportion of correctly identified

model specifications via a min *AIC* selection is a straightforward endeavor. Correct

model selection percentages, however, only represent the lower bound of correct

empirical evidence percentage for each of the hypotheses. Selection of an incorrect

model can, depending on the hypothesis, still provide the correct hypothesis conclusion.

Table 5 shows model selections corresponding to correct empirical evidence regarding a

particular hypothesis for different population conditions.

Table 5.
Correct Empirical Evidence by Model Selection (Normal Distribution)

Population Conditions*		Correct Empirical Evidence			
ϕ	σ	Correct	for ϕ	for μ	for σ^2 **
$\phi = 0$	$\sigma_C \neq \sigma_T$	Model #1	Model #1, 2, 3, 4	Model #1, 2, 5, 6	Model #1, 3, 5
	$\sigma_C = \sigma_T$	Model #2	Model #1, 2, 3, 4	Model #1, 2, 5, 6	Model #2, 4, 6
$0 < \phi < 1$	$\sigma_C \neq \sigma_T$	Model #5	Model #5, 6	Model #1, 2, 5, 6	Model #1, 3, 5
	$\sigma_C = \sigma_T$	Model #6	Model #5, 6	Model #1, 2, 5, 6	Model #2, 4, 6

* Note: All population conditions operated with a population treatment effect.
** Note: Evidence for this condition is only possible when either the control or treatment population is
posited from a normal distribution. Chapter 4 will provide another distributional example.

For example, consider population conditions which had a treatment effect, equal

population variances, and treatment nonresponse (Model #6), but the min *AIC* selection

was Model #5. Despite being the incorrect model representation, the treatment effect and

mixture hypotheses conclusions are still correct. While this process does not make

probabilistic inferences, the complement to proportions of correct hypotheses conclusions

is analogous to Type I error.   The culmination of the empirical study is a series of tables,

Appendices 2-5, providing successful hypotheses conclusions as a function of the

controlled parameters for reference in future applied research. Of course, a min *AIC*

selection does not guarantee quality of fit, at worst resulting in the least poor choice among ill-fitting models. To help prevent this possibility, pre and post-model selection tests supporting a researcher's parametric distribution choices are presented in Chapter 4.

Finally, convergence can be problematic in a comparative model selection, particularly if the optimization algorithm of the *correctly* specified model fails to converge. Enumerated in the previous section, convergence is a function of many elements: data, model specification, starting values, number of iterations, number of attempts, and model fit conditions. Convergence rates are retained for the three models requiring an optimization algorithm, Models #3, #5, and #6, with particular focus on the mixture model specifications.

2.3.2 Population Level

The recapturing of population level parameters is based on the *AIC* selected model where the overall composition of correct and incorrect models for a series of trials differs for every experimental condition. *Bias* and mean squared error *(MSE)* will be retained for $\hat{\phi}$. The general form for *bias* is

$$bias_{\hat{\theta}} = bias(\hat{\theta}, \theta) = E(\hat{\theta}) - \theta \tag{69}$$

where we will use the average sample results as an estimate of the expected value

$$\bar{\hat{\theta}} = \frac{1}{\#\,\text{trials}} \sum_{i=1}^{\#\,\text{trials}} \hat{\theta}_i \tag{70}$$

and *MSE* is

$$MSE(\hat{\theta}) = V(\hat{\theta}) + bias^2(\hat{\theta}, \theta) \tag{71}$$

where the expected variance is estimated from the sample results

$$\frac{\sum_{i=1}^{\#\text{trials}} \left(\hat{\theta}_i - \bar{\hat{\theta}} \right)^2}{\#\text{trials}} \tag{72}$$

with 500 trials per experimental condition. Applying these general formulas, the estimated *bias* and *MSE* for $\hat{\phi}$ are

$$\frac{1}{500} \sum_{i=1}^{500} \left(\hat{\phi}_i - \phi \right) \tag{73}$$

and

$$\frac{1}{500} \sum_{i=1}^{500} \left(\hat{\phi}_i - \bar{\hat{\phi}} \right)^2 - bias^2{}_{\hat{\phi}} \tag{74}$$

respectively. In lieu of providing *bias* and *MSE* values for each distributional parameter estimate, because treatment effects are commonly reported, either $\hat{d}_u$ or $\hat{d}^*$, these composite estimates are evaluated instead. Despite both being representative of a treatment effect, recapturing of population information may differ, where $\hat{d}_u$'s *bias* and *MSE* are estimated as

$$\left(\frac{1}{500} \sum_{i=1}^{500} \left(|\hat{\mu}_{Ci} - \hat{\mu}_{Ti}| \right) \right) - \left(|\mu_C - \mu_T| \right) \tag{75}$$

and

$$\frac{1}{500} \sum_{i=1}^{500} \left(\left(| \hat{\mu}_{Ci} - \hat{\mu}_{Ti} | \right) - \bar{d}_u \right)^2 + bias^2{}_{\hat{d}_u} \tag{76}$$

respectively. *Bias* and *MSE* for the standardized treatment effect estimate, $\hat{d}^*$, are calculated using

$$\left(\frac{1}{500} \sum_{i=1}^{500} \frac{|\hat{\mu}_{Ci} - \hat{\mu}_{Ti}|}{\sqrt{\dfrac{\hat{n}^*_{Ci}\hat{\sigma}^2_{Ci} + \hat{n}^*_{Ti}\hat{\sigma}^2_{Ti}}{N}}} \right) - \frac{|\mu_C - \mu_T|}{\sqrt{\dfrac{n^*_C\sigma^2_C + n^*_C\sigma^2_T}{N}}} \tag{77}$$

and

$$\frac{1}{500} \sum_{i=1}^{500} \left(\frac{|\hat{\mu}_{Ci} - \hat{\mu}_{Ti}|}{\sqrt{\dfrac{\hat{n}^*_{Ci}\hat{\sigma}^2_{Ci} + \hat{n}^*_{Ti}\hat{\sigma}^2_{Ti}}{N}}} - \overline{\hat{d}}^{\,*} \right)^2 + bias^2_{\hat{d}*} \tag{78}$$

Relating these population level measures to model level measures in the last section as well as highlighting the variety of outcomes, the following results are possible for a single trial:

a. Correct model selection can result in treatment effect estimates, $\{\hat{d}^*, \hat{d}_u\}$ which greatly deviate from their true value. Further, inaccuracies between these estimates can be varied.

b. Correct model selection can result in a mixing proportion estimate, $\hat{\phi}$, which greatly deviates from its true value.

c. An incorrect model selection can result in exact estimates for either treatment effect representation.

d. An incorrect model selection, dependent on population conditions, can produce an exact estimate of the mixing proportion. For example, if model #5 was the correct specification, a model #6 selection can result in an exact estimate of the mixing proportion.

To mitigate sampling error effects as well as model convergence issues, 500 trials were conducted providing well supported proportions of hypotheses conclusions.

50

Despite differences in model selection composition, the following results are possible for an experimental condition:

    a. A 100% correct model selection rate can result in a positive or negative *bias* for one or both of the treatment effect estimates.

    b. A large proportion of incorrect model selection can result in unbiased treatment effect estimates.

    c. A 100% correct model selection rate can result in positive or negative *bias* of the mixing proportion estimate. For experimental conditions where the population mixing proportion is 0, the *bias* can only be positive.

    d. A large proportion of incorrect model selection can result in an unbiased mixing proportion estimate. When the population mixing proportion is 0, however, model selection must be entirely composed of Models #1-4 in order to obtain an unbiased result.

Comparative performances in regards to *bias* and *MSE* between the treatment effect representation estimates are not tenable because of their different scales.

    2.3.3 Individual Level

    Invariably accompanying the supposition of nonresponse to a particular treatment is an interest in class membership of each individual in the treatment sample. With the selected model's parameter estimates, Bayes' theorem is utilized post-model selection, estimating the probability of being a treatment nonrespondent for each jth individual as

$$\hat{\pi}_j = \frac{\hat{f}_{\mathrm{C}}(x_{\mathrm{T}j}) * \hat{\phi}}{\hat{f}_{\mathrm{C}}(x_{\mathrm{T}j}) * \hat{\phi} + \hat{f}_{\mathrm{T}}(x_{\mathrm{T}j}) * (1 - \hat{\phi})} \quad \text{for } j = 1, 2, ..., n_{\mathrm{T}} \tag{79}$$

While z^{*} requires a population mixing proportion in its calculation, individual probabilities are not required, where the relationship from individual to population level can be expressed as

$$\hat{\phi} = \frac{\displaystyle\sum_{j=1}^{n_{\mathrm{T}}} \hat{\pi}_j}{n_{\mathrm{T}}} \tag{80}$$

Measuring the recovery of group membership for a series of trials can be accomplished two ways, regardless of population conditions and subsequent model selection. The first method provides an average error per treatment sample respondent between their predicted posterior probability and known membership per experimental condition

$$\overline{\pi}_{error} = \frac{1}{500 n_{\mathrm{T}}} \sum_{i=1}^{500} \sqrt{\sum_{j=1}^{n_{\mathrm{T}}} (\hat{\pi}_{j,i} - \pi_j)^2} \tag{81}$$

where $\hat{\pi}_{j,i}$ is estimated post model selection per trial i using Equation (79) and π_j values are known as part of the data generation process described in section 2.1, where the probability of being a treatment nonrespondent is either 0 or 1. Measurement per individual facilitates comparisons across different treatment sample sizes.

A second measure, percentage of correct classification, while coarser, is more easily interpreted. For each individual j in trial i, perform the following calculation

$$\text{if } \hat{\pi}_{j,i} \leq .50, \text{ then } \hat{\pi}_{j,i}^{*} = 0; \text{ else } \hat{\pi}_{j,i}^{*} = 1$$
$$\text{for } j = 1, 2, ..., n_{\mathrm{T}} \quad \text{for } i = 1, 2, ..., 500 \tag{82}$$

where the average overall correct classification percentage becomes

$$\bar{\pi}_{\%\text{Class}} = 1 - \frac{1}{500 n_T} \sum_{i=1}^{500} \sum_{j=1}^{n_T} (\hat{\pi}_{j,i}^* - \pi_j)^2 \qquad (83)$$

Measurement as an overall classification as well facilitates comparisons across different

treatment sample sizes. A single trial Model #1-4 selection is not very interesting with

the predicted nonresponse membership of each treatment sample respondent being 0.

More interesting is the case where model selection provides empirical evidence

supporting nonresponse. Consider Figure 6 as an example where a mixture model was an

incorrect selection when population conditions had zero treatment nonresponse.

Figure 6.
Graph: Single Trial Illustration of Posterior Probabilities at $\phi = 0$, $n_T = 60$,
      Incorrect Model Selection
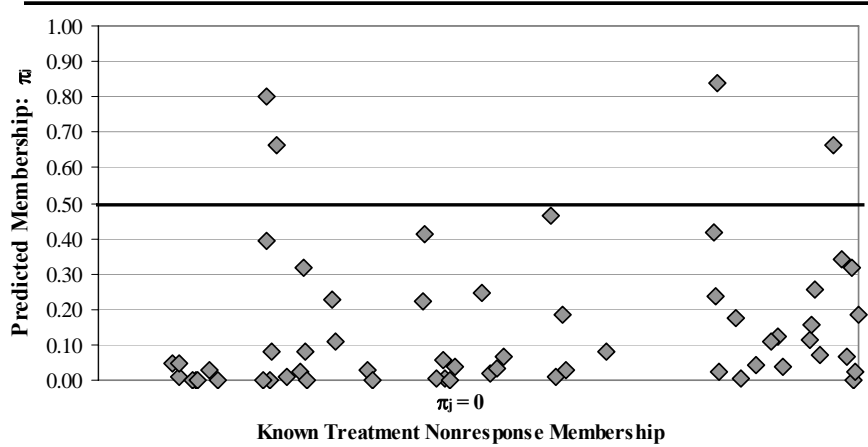
Respondents above the 0.50 threshold will be incorrectly classified using the coarse

assessment measure presented. Continuing this example, the average respondent error per

treatment respondent is 12.6%, while the correct classification percentage is 93.3%.

Another possible outcome is Figure 7, where the correct model is selected and treatment

nonresponse exists.

Figure 7.
Graph: Single Trial Illustration of Posterior Probabilities at $\phi = .20$, $n_T = 60$,
Correct Model Selection
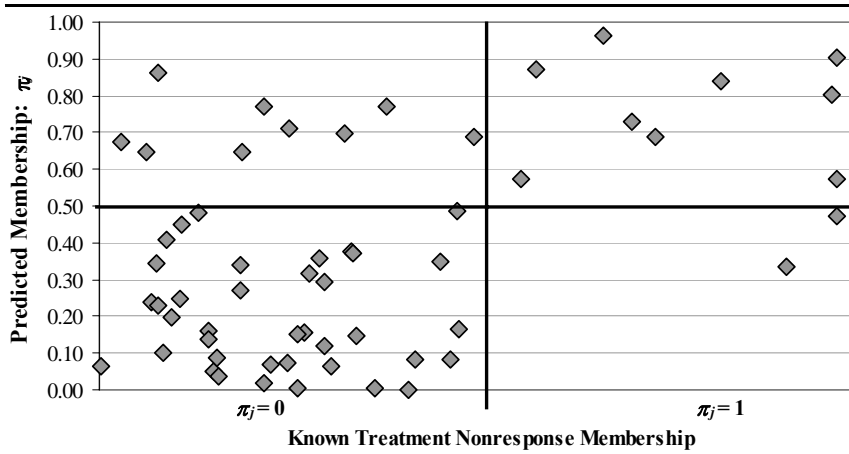
Respondents in the upper left and lower right quadrants are incorrectly classified, where

this trial had a 25.2% average probability error per respondent and an overall

classification rate of 81.7%. Similar to the scenarios possible between the model and

population level measures, there is no guarantee performances of individual measures

coincide with performance at either the model or population level.

2.4 Simulation Parameter Validation (Pilot Results)

This section validates, through a pilot study, that the three study parameters

chosen for manipulation, $\{z^*, \phi, n_T\}$, effect changes in selected performance measures,

focusing on correct model selection and mixture hypothesis evidence. Population means

values are fixed with a treatment effect where

$$\mu_C = 20$$
$$\mu_T = 30$$

(84)

and sample sizes, standard deviations, and the mixing proportion adjusted to differentiate

experimental conditions. Using Equation (84), the unstandardized treatment effect

remains constant at

$$d_u = |\mu_C - \mu_T| = 10 \tag{85}$$

where d^* will change as part of other experimental conditions. A second equally useful

purpose from the pilot study provides information on effective boundaries for the

controlled parameters. Theoretical analysis from section 2.2 helped only to identify these

parameters, not describe the shape of their relationships to these model level measures.

Inspection of these relationships can eliminate conditions where no change in model

selection occurs with greater magnification in ranges of change in the comprehensive

study. Beyond the systematic variation of the study parameters is inclusion of other

design characteristics, shown in the four configurations varying sample size and variance

ratios presented in section 2.2.1, represented as *a* and *b* from Equations (58) and (59).

2.4.1 False Mixture Classification with Zero Treatment Nonresponse

The easiest experimental conditions to explore occur where

$$\phi = 0 \tag{86}$$

and the effective treatment group sample size becomes

$$n_T^* = n_T \tag{87}$$

The last section related the proportion of incorrect hypotheses conclusions from min *AIC*

selection as a proxy to Type I error in inferential tests. Continuing this analogy, Type II

error, false acceptance of a mixture model, is an equally bad result and a necessary

component to conduct power analysis presented in Chapter 4.

Returning to the experimental conditions manipulated to obtain the z^* values in the pilot study, additional restrictions were imposed such that

$$50 \leq n_C, n_T \leq 400$$

$$n_C \leq n_T \tag{88}$$

$$\sigma_C \leq \sigma_T$$

where n_T values were varied in z^* value construction. In the comprehensive study, n_T is fixed to delineate the changes in model selection due solely to z^*. Sample sizes chosen are consistent with small to large scale studies, while sample size and variance ratios ranged from

$$a = \frac{n_C}{n_T} = (0.66, 1.00)$$

$$\tag{89}$$

$$b = \sigma_{rat}^2 = \frac{\sigma_C^2}{\sigma_T^2} = (0.75, 1.00)$$

where values of 1.00 indicate equality. This range of ratios represents comparatively larger treatment sample sizes and treatment population variances. Ratios greater than 1.00 will be incorporated into the comprehensive study contingent on the pilot results. Table 6 provides the population conditions, a d^* value, and experimental design type (sample size and variance ratio configurations) used in the 36 sets of conditions, each having 500 trials. For all experimental condition tables in the dissertation, population standard deviations are reported in lieu of variances retaining the same metric as population means to improve interpretation.

Table 6.
Empirical Conditions for various z^* when $\phi = 0$, $\sigma_C = \sigma_T$ and $\sigma_C < \sigma_T$

					z^*				
	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5
Type	1	2	1	2	3	4	1	1	2
d^*	0.28	0.48	0.62	0.90	1.00	1.25	0.98	1.10	0.95
σ_C	32.00	17.00	13.00	10.00	10.00	8.00	9.00	8.00	9.90
σ_T	38.00	24.00	18.00	12.10	10.00	8.00	11.00	10.00	11.10
n_C	50	54	50	50	50	54	100	110	200
n_T	60	54	68	50	77	54	125	125	200
Type	3	4	3	4	1	2	3	4	3
d^*	0.29	0.45	0.67	0.90	1.03	1.24	1.01	1.22	1.03
σ_C	35.00	22.00	15.00	11.10	9.00	7.00	9.90	8.20	9.70
σ_T	35.00	22.00	15.00	11.10	10.30	9.00	9.90	8.20	9.70
n_C	50	60	50	50	50	55	100	97	150
n_T	62	60	61	50	63	55	123	97	196

					z^*				
	11	13	15	17	19	22	26	30	34
Type	4	3	2	4	3	1	2	4	1
d^*	0.91	1.25	1.11	1.45	1.64	1.63	2.59	2.27	2.68
σ_C	11.00	8.00	8.00	6.90	6.10	5.30	3.50	4.40	3.50
σ_T	11.00	8.00	9.90	6.90	6.10	6.70	4.20	4.40	3.90
n_C	293	204	365	275	250	300	202	349	275
n_T	293	230	365	275	290	397	202	349	363
Type	1	2	4	1	2	3	4	2	3
d^*	1.01	1.30	1.22	1.50	1.66	1.67	2.44	2.36	2.70
σ_C	9.00	7.00	8.20	6.20	5.50	6.00	4.10	3.70	3.70
σ_T	10.50	8.30	8.20	7.00	6.50	6.00	4.10	4.70	3.70
n_C	200	199	303	225	262	320	227	322	290
n_T	262	199	303	280	262	382	227	322	348

Graphical results summarizing the correct mixture hypothesis conclusion in Figure 8 are extremely informative. A definitive relationship is evident between z^* and correct empirical evidence of a mixture, with surprising clarity given variations in treatment sample sizes.

Figure 8.
Graph: Correct Mixture Hypothesis Conclusion when $\phi = 0.0$
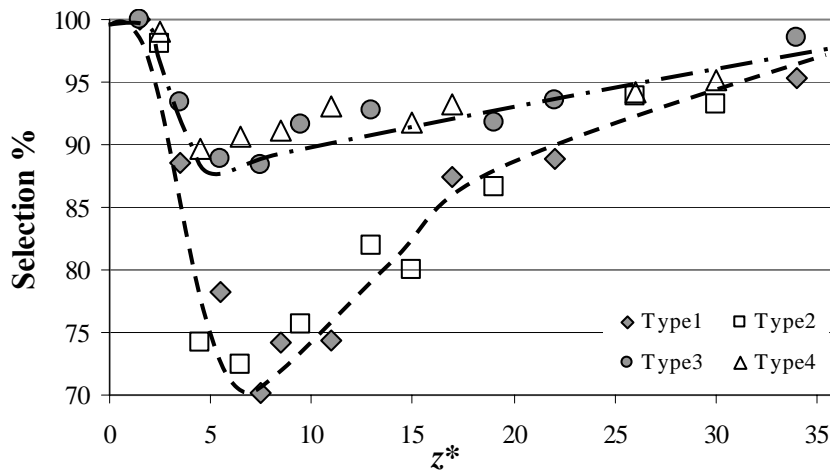
The results are more informative than confirmation of z^*'s relation to the correct mixture hypothesis conclusion:

a. The behavior of models with similar variance ratios, $\{b = 1, b \neq 1\}$, is quite consistent irrespective of sample size ratios, represented as two pairs, {Type 3, Type 4} and {Type 1, Type 2}, which reflect design characteristics from section 2.2.1 in terms of sample size and variance ratios. This does not imply sample size differences are unimportant, instead that differences are subsumed within z^*.

b. A clear distinction exists in performance when variance ratios differ where Type 3 and Type 4 correspond to a ratio of 1.0 and equality. Performance results by variance ratio clearly indicate these differences are not subsumed in the z^* measure, unlike the sample size ratio.

c. The interest from these conditions is the false selection of any mixture model which is the complement of the correct mixture hypothesis conclusion

$$\beta^*_{Mix} = 1 - \%\text{Correct}_{Mix} \mid \phi = 0, z^*, \sigma^2_{rat} = ? \tag{90}$$

As an example, consider the case of a zero treatment nonresponse with a z^* value of 5. The false selection percentage of a mixture model with equal population variances is 11%, where inequalities in the range of ratios evaluated is 27%.

The other model level measure, correct model selection, is shown in Figure 9 with the 36 experimental conditions equally divided between Model #1 and #2 as the correct model specification.

Figure 9.
Graph: Correct Model Selection when $\phi = 0.0$
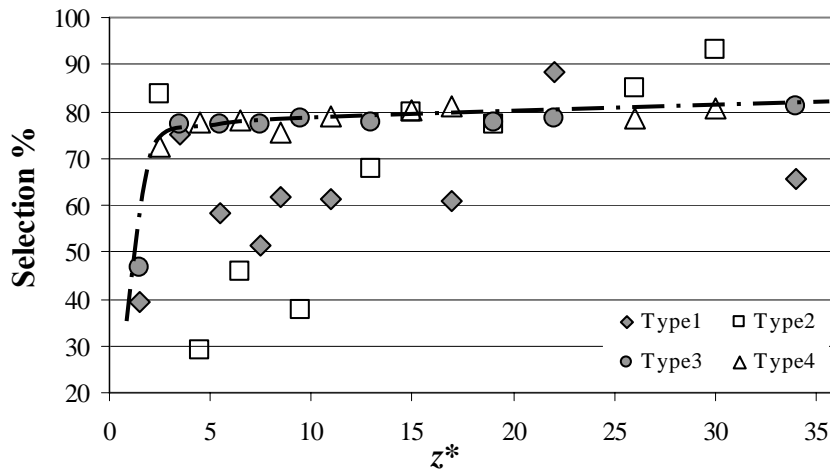
A disparity in results between variance ratios with this measure is reaffirmed and even more pronounced. Correct model selection as a function of z^* was consistent irrespective of sample size ratios when σ^2_{rat} was 1.0, which was not observed with unequal variances. These findings coincide with another multiple sample empirical study not involving mixtures where heterogeneity in group variances detrimentally affected correct model identification rates (Huang, 1995). With equal variance conditions, a sharp decline in correct model selection occurred as z^* dropped below 4, where model selection increasingly became Model #3 or Model #4 concluding no treatment effect.

Based on these results, construction of remaining pilot study conditions was simplified to validate mixing proportion and treatment sample size impact on the selected performance measures. First, because performance appears conditioned on the variance ratio, the remainder of the pilot study considers only variance equality. Second, because no performance differences were observed due to changes in the sample size ratio, sample sizes were made equal.

2.4.2 Impact of Mixing Proportion

To illustrate model selection performance impact due to the mixing proportion, conditions were established fixing both z^* and n_T. Table 7 provides the experimental conditions, including d^*, for six different mixing proportions and two n_T values with the same fixed z^*.

Table 7.
Empirical Conditions for various ϕ, z^* fixed at 15, Type 4

| | $n_T = 100$ | | | | | | $n_T = 200$ | | | | | |
| | ϕ | | | | | | ϕ | | | | | |
	0.02	0.05	0.10	0.20	0.35	0.50	0.02	0.05	0.10	0.20	0.35	0.50
d^*	2.12	2.12	2.13	2.16	2.26	2.45	1.50	1.50	1.51	1.53	1.60	1.73
$\sigma_C,\ \sigma_T$	4.71	4.71	4.69	4.62	4.42	4.08	6.67	6.66	6.63	6.53	6.24	5.77

Figures 10 and 11 present the findings by n_T condition, with each graph displaying correct model selection and correct mixture hypothesis conclusion percentages.

Figure 10.

Graph: Model Selection, $\sigma^2_{rat} = 1.0$, $z^* = 15$, $n_T = 100$
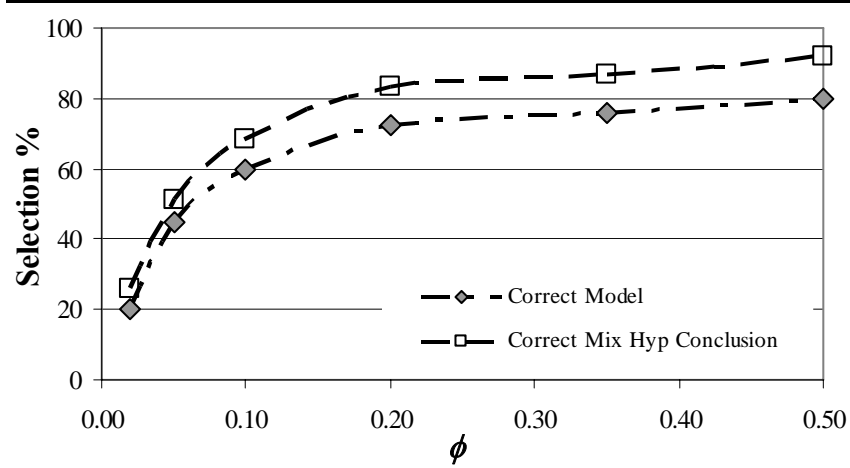
Figure 11.

Graph: Model Selection, $\sigma^2_{rat} = 1.0$, $z^* = 15$, $n_T = 200$
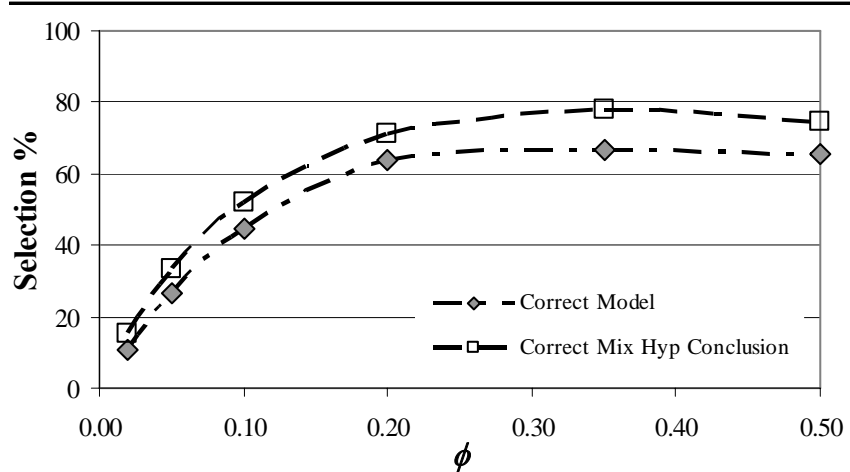
These figures confirm both selected performance measures are affected by the mixing proportion, improving as ϕ deviates from 0. A comparison of the plots reveals a counterintuitive result; larger sample sizes have worse correct model and hypothesis conclusion rates. Recall, however, z^* is not independent of sample size, so maintaining a fixed z^* value with larger sample sizes requires increased standard deviations.

      A commonly expected result regarding increased sample size can be observed using d^* as the treatment effect measure instead of z^*, where d^* is independent of sample

size. Using the results from sample sizes of 100, another set of trials was conducted with

the same standard deviation values while increasing the sample sizes each to 250. While

the d^* values remained the same, this increased z^* to over 21. The results, presented in

Figure 12, which co-label ϕ and d^* on the x-axis, indicate improvement in correct

mixture hypothesis conclusion with larger samples.

Figure 12.
Graph: Correct Mixture Hypothesis Conclusion, $\sigma^2_{rat} = 1.0$,
with d^* at $z^* = 15$ conditions
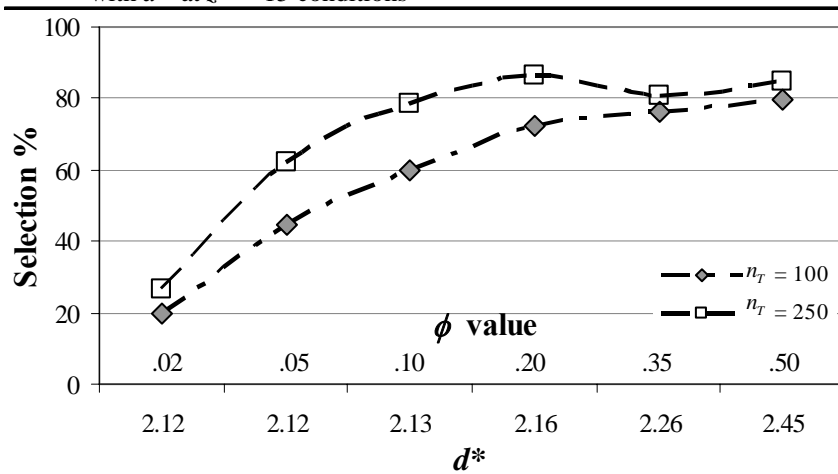
2.4.3 Impact of Treatment Effect

To verify the impact of z^* on model level performance with conditions of

treatment nonresponse, ϕ and n_T values were fixed. Three different treatment effect

values and four different n_T values with a constant ϕ of 0.30, presented in Table 8, were

evaluated.

Table 8.
Empirical Conditions for various n_T, ϕ fixed at 0.30, Type 4

	$z*$ fixed at 11				$z*$ fixed at 15				$z*$ fixed at 19			
	n_T				n_T				n_T			
	34	50	67	100	34	50	67	100	34	50	67	100
$d*$	2.79	2.30	1.99	1.63	3.80	3.14	2.72	2.22	4.83	3.98	3.44	2.28
$\sigma_C,\ \sigma_T$	3.58	4.34	5.02	6.13	2.63	3.18	3.68	4.50	2.07	2.51	2.91	3.55

Figure 13 illustrates the correct mixture hypothesis conclusion while Figure 14 illustrates

the correct model selection.

Figure 13.
Graph: Correct Mixture Hypothesis Conclusion, $\sigma^2_{rat} = 1.0$, $\phi = 0.30$
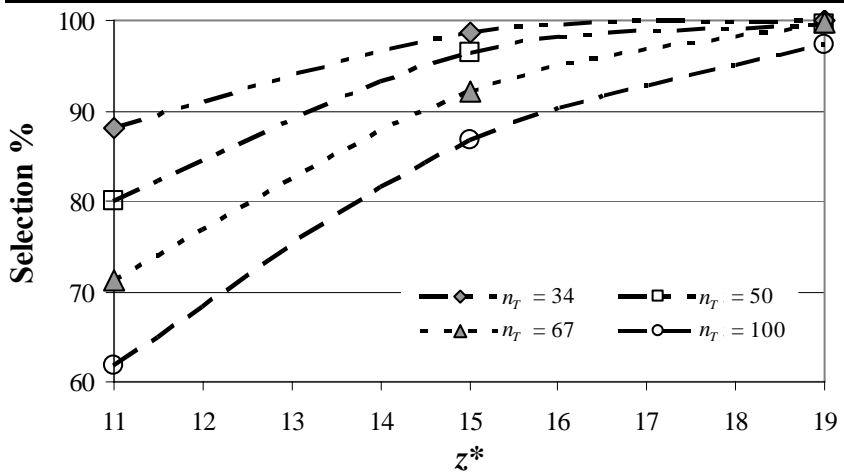
Figure 14.
Graph: Correct Model Selection, $\sigma^2_{rat} = 1.0$, $\phi = 0.30$
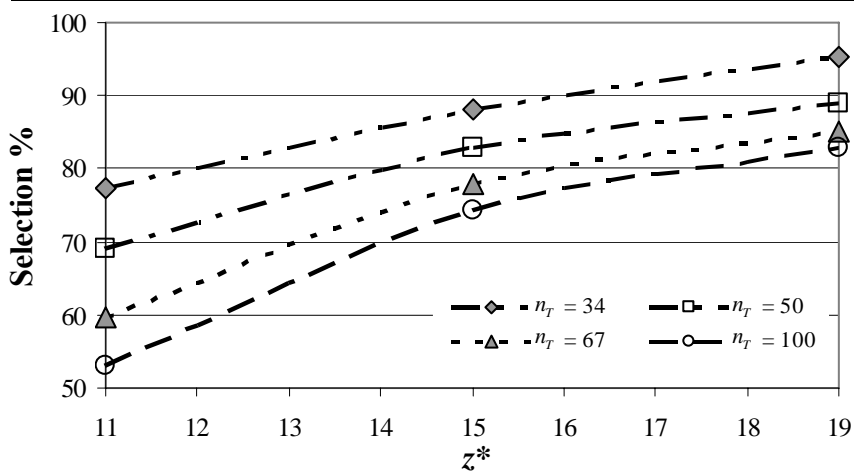
Larger z^* values improved both model level performance measures, where noted

previously, results were comparatively worse with larger n_T values when z^* was fixed. A

mixing proportion of 0.10, reduced from 0.30, with significantly larger n_T values was also

evaluated, with experimental conditions provided in Table 9 and graphical results in

Figures 15 and 16.

Table 9.
Empirical Conditions for various n_T, ϕ fixed at 0.10, Type 4

	z^* fixed at 11				z^* fixed at 15				z^* fixed at 19			
	n_T				n_T				n_T			
	100	150	200	300	100	150	200	300	100	150	200	300
d^*	1.56	1.28	1.10	0.90	2.13	1.74	1.51	1.23	2.70	2.20	1.91	1.56
σ_C, σ_T	6.40	7.83	9.05	11.10	4.69	5.75	6.63	8.12	3.70	4.54	5.23	6.41

Figure 15.
Graph: Correct Mixture Hypothesis Conclusion, $\sigma^2_{rat} = 1.0$, $\phi = 0.10$

64
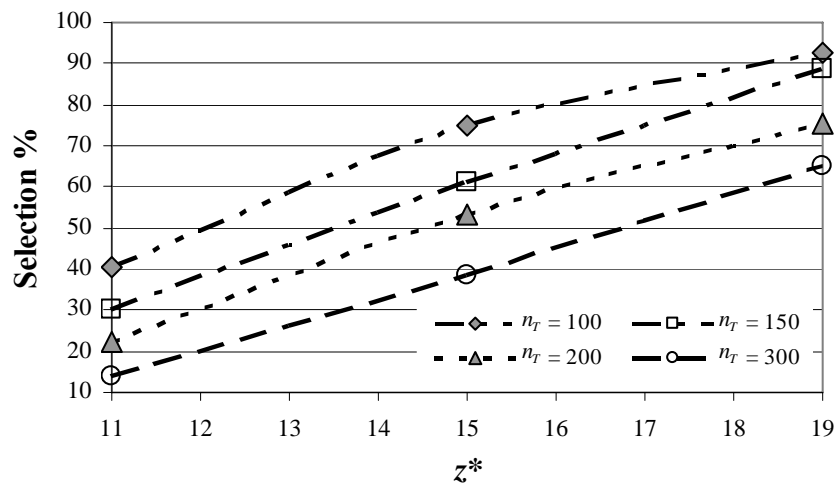
Figure 16.

Graph: Correct Model Selection, $\sigma^2_{rat} = 1.0$, $\phi = 0.10$

Selection % (y-axis: 0, 10, 20, 30, 40, 50, 60, 70, 80, 90)

z^* (x-axis: 11, 12, 13, 14, 15, 16, 17, 18, 19)

Legend: $n_T = 100$; $n_T = 150$; $n_T = 200$; $n_T = 300$
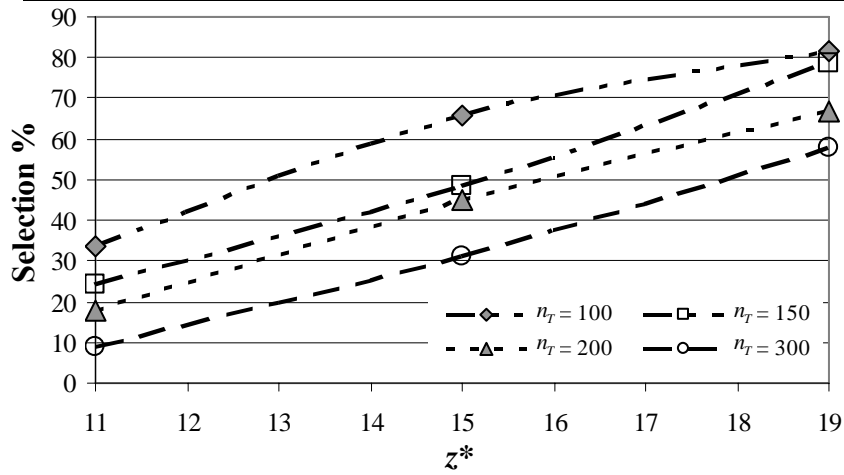
The results are consistent with those of the 0.30 mixing proportion previously evaluated. Comparatively, however, the smaller mixing proportion had much lower correct hypothesis conclusion and model selection rates, even with larger n_T values, with the greatest disparity occurring at smaller z^* values.

2.4.4 Impact of Treatment Group Size

While the relationship between n_T and model selection was introduced last section, this section extends the previous study considering larger n_T values. Extension of Tables 8 and 9 increasing the total number of n_T conditions to six are provided for both mixing proportions in Table 10.

Table 10.

Empirical Conditions for various z^*, various ϕ, Type 4

	$\phi = 0.30$						$\phi = 0.10$					
	$n_T = 133$			$n_T = 167$			$n_T = 400$			$n_T = 500$		
	z^*			z^*			z^*			z^*		
	11	15	19	11	15	19	11	15	19	11	15	19
d^*	1.41	1.93	2.44	1.26	1.72	2.18	0.78	1.07	1.35	0.70	0.95	1.21
σ_C, σ_T	7.07	5.19	4.09	7.93	5.81	4.59	12.79	9.38	7.41	14.30	10.49	8.28

Because this is an extension of results presented in previous section, d^* was used on the

x-axis instead. These relationships are consistent with those presented in section 2.4.2

illustrating dramatic changes in performance within a relatively small range of d^* values

shown in Figures 17 and 18.

Figure 17.
Graph: Correct Mixture Hypothesis Conclusion, $\sigma^2_{rat} = 1.0$ for $\phi = 0.30$ and $\phi = 0.10$ with d^*
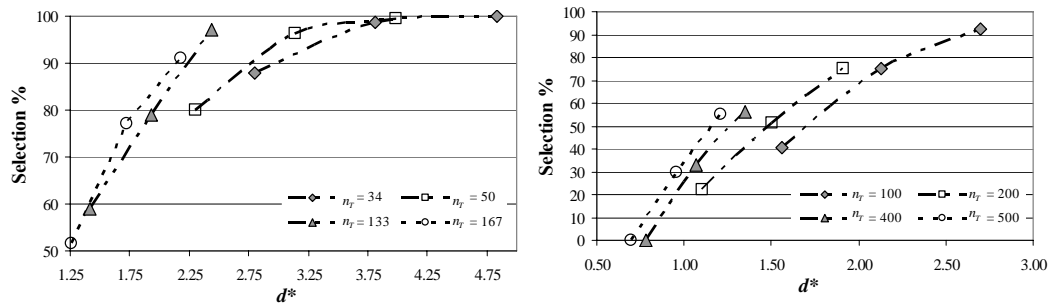
Figure 18.
Graph: Correct Model Selection, $\sigma^2_{rat} = 1.0$ for $\phi = 0.30$ and $\phi = 0.10$ with d^*
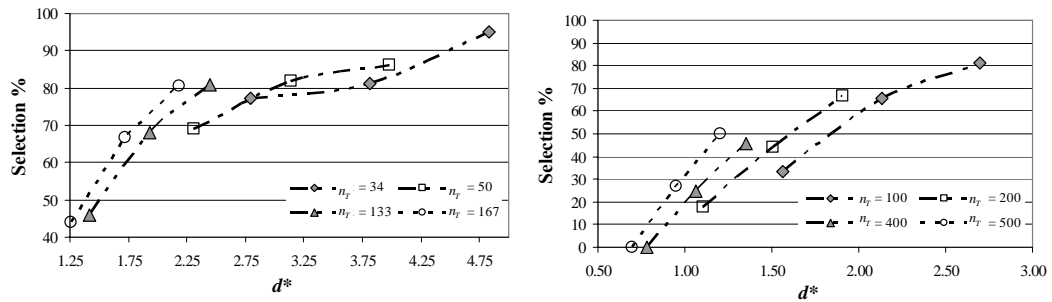
Correct model selection and mixture hypothesis conclusions remain comparatively lower

for smaller mixing proportions.

2.4.5 Treatment Effect Relationships

Selection of z^* as a controlled parameter, despite validation from a pilot study,

makes interpretation of some results initially confusing. Because d^* also demonstrated

relationships with the selected model level performance measures, connections can be

explored between these treatment effect representations. Such exploration can also explain other results observed in the pilot study.

For instance, recall from section 2.4.1 the invariance in model selection in varied sample size ratios resulting in the ratio subsequently fixed to 1.0. When treatment nonresponse is present, the sample size ratio, a, is adjusted using effective sample sizes as

$$a* = \frac{n_C^*}{n_T^*} \tag{91}$$

With zero treatment nonresponse, this reduces to Equation (58), otherwise with equal sample sizes, the effective ratio becomes

$$a* = \frac{(1+\phi)}{(1-\phi)} \tag{92}$$

The maximum z^* value, $z^*_{\max}$, with a σ^2_{rat} of 1.0 occurs when a^* is 1, where other treatment effect representations, d^* and d_u, are unaffected by changes in a^*. Accordingly, if planning an experimental design supportive of unequal samples, treatment sample size should be increased to account for nonresponse. For studies constrained to equal sample sizes, z^* will be less than its maximum depending on the amount of nonresponse. The deviation from its maximum, or penalty, independent of sample size, is characterized as

$$z^*_{\text{adj}} = z^*_{\max} \sqrt{(1-\phi^2)} \tag{93}$$

with a graphical illustration in Figure 19.

Figure 19.
Graph: Percentage of $z*_{max}$ over ϕ; with Equal σ^2 and Sample Sizes
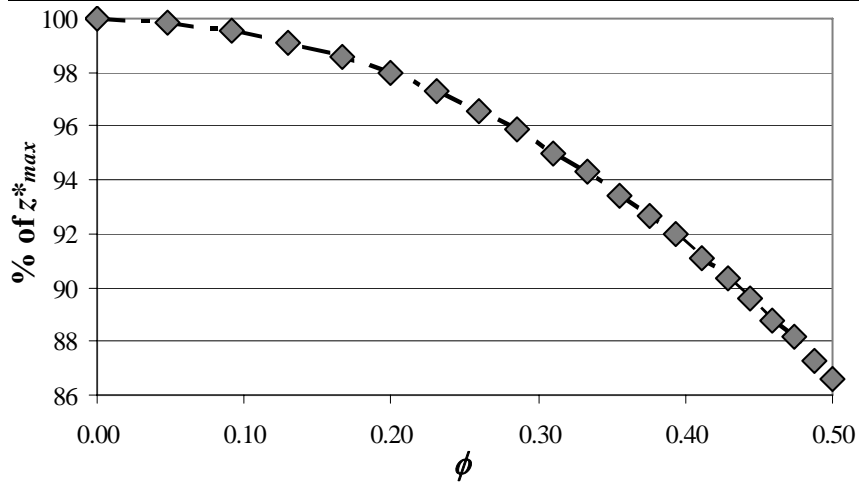
Adjusting the information from this figure to experimental conditions in section 2.4.1, sample size ratios evaluated corresponded to less than a 4% change from the maximum z^* value resulting in no observable change in model level performance. This suggests z^* is robust to moderate effective ratio differences as a result of either experimental design or treatment nonresponse in equal samples.

Relationships between treatment effect representations can be developed for the pilot study case of equal variances, substituting relationships from Equation (62) into Equation (63) produces

$$z^* = \frac{d^* \sigma}{\sqrt{\dfrac{\sigma^2}{n_C^*} + \dfrac{\sigma^2}{n_T^*}}} \tag{94}$$

where after some algebraic manipulation

$$z^* = d^* \sqrt{\frac{(1-\phi^2)n_C}{2}} \quad \text{for } n_C = n_T \tag{95}$$

and a general representation which accommodates unequal sample sizes

$$z^* = d^* \sqrt{\frac{n_C^* \, n_T^*}{N}} \qquad (96)$$

Formulaic relationships between d^* and z^* are more complicated with the unequal

population variances evaluated in the comprehensive study. While d^* remains

independent of sample sizes, it is no longer independent of ϕ, σ^2_{rat}, and sample size ratio

as shown later in section 3.2.4.

2.5 Empirical Conditions

The pilot study successfully demonstrated the parameters selected for systematic

manipulation influence correct model selection and the mixture hypothesis conclusion

results. These same parameters are used in the comprehensive study of normal

population distributions including variance ratios other than 1.0. The pilot study was also

successful in identifying boundaries for z^*. Further, as a means of limiting the total

number of experimental conditions while taking advantage of model level performance

relation to d^*, an empirical condition will not be evaluated if

$$0.25 > d^* > 4.0 \qquad (97)$$

This coincides with the pilot study's region of greatest change, where violations are

annotated within the experimental condition tables in Appendix 1. Sample size ratios

remain fixed at 1.0 while population means of 20 and 30 for the control and treatment

populations are unchanged. Unlike the pilot study, this multifactorial study is concerned

with not only select model level performance measures, but all performance criteria

presented in section 2.3. The schematic of the comprehensive study presented in Figure

20 consists of over 580 experimental conditions, with added specificity in the following

subsections.

Figure 20.
Construct for a Comprehensive Empirical Study

Normal Distribution
$$(\sigma^2_{rat} = \sigma^2_C / \sigma^2_T = 1.00)$$
Treatment Conditions
z^*: 9 Values ϕ: 6 Values n_T: 4 Values

Normal Distribution
$\sigma^2_{rat} = 0.50$

z^* : 9 Values
ϕ : 6 Values
n_T : 3 Values

Normal Distribution
$\sigma^2_{rat} = 2.00$

z^* : 9 Values
ϕ : 6 Values
n_T : 3 Values

Normal Distribution
$\sigma^2_{rat} = 0.33$

z^* : 9 Values
ϕ : 4 Values
n_T : 2 Values

Normal Distribution
$\sigma^2_{rat} = 3.00$

z^* : 9 Values
ϕ : 4 Values
n_T : 2 Values
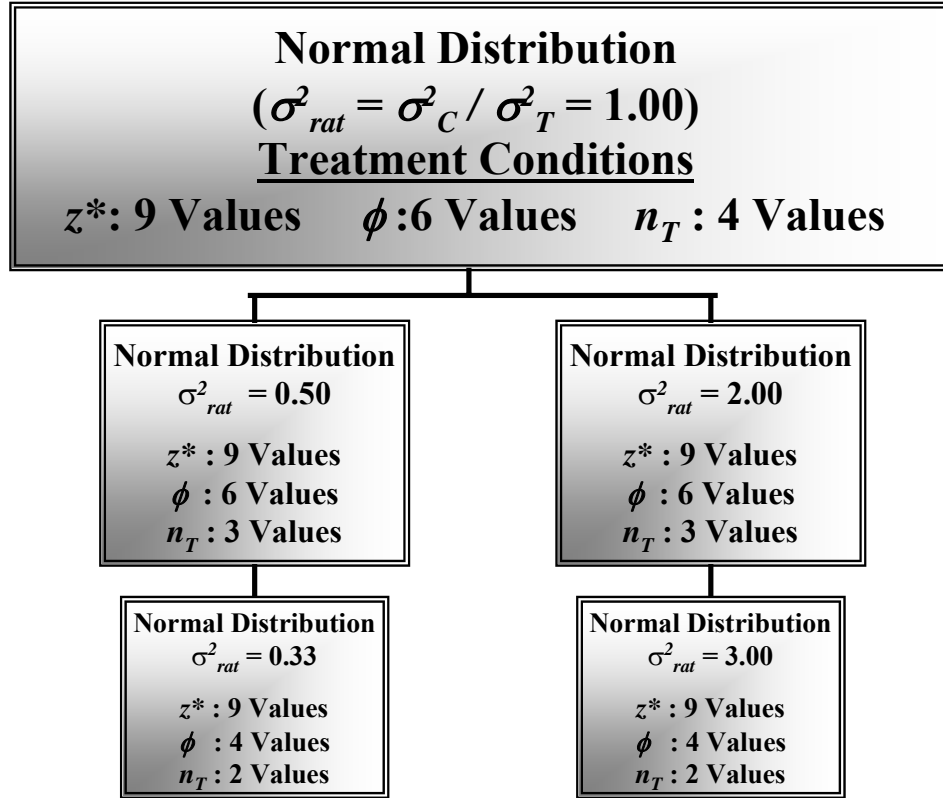
2.5.1 Normal Distributions with Variance Equality:

This portion of the study provides a more thorough exploration of results

observed in the pilot study, where the z^* values were

$$\{3, 4, 7, 10, 15, 20, 25, 30, 40\} \tag{98}$$

coinciding with the region of greatest changes in the pilot study. The population mixing

proportion was varied

$$\{0, 0.05, 0.10, 0.20, 0.35, 0.50\} \tag{99}$$

Selection of a zero mixing proportion illuminates the issue of false classification, treatment sample sizes were fixed at

$$\{50, 100, 200, 350\} \qquad (100)$$

Conditions of zero treatment nonresponse represent Model #2, while Model #6 is the correct model specification otherwise.

2.5.2. Normal Distributions with Unequal Variances

While the normal distribution's great flexibility to effectively characterize observed and transformed data structures make it the predominant choice as a parametric distribution, its selection has some negative implications. First, under this methodology, it requires evaluation of more model representations to obtain empirical evidence, twice as many as a single parameter distribution presented in Chapter 4. Second, generalization of any results for data adhering to normal distributions is challenging, highlighted in the simple pilot study; normal distributions with unequal population variances performed differently on the selected measures. Such differences likely extend to the additional performance measures presented in section 2.3. Further exacerbating this challenge, the pilot study only considered variance ratios less than 1.0, yet ratios greater than 1.0 in applied settings are just as common. Theoretical analysis of the likelihood function offered no indication this methodology is symmetric in performance for variance ratios equidistant from 1.0 (e.g. 0.50 and 2.0). Finally, there is an issue of selecting appropriate variance ratios which illustrate a substantive change in methodological performance. Zimmerman (2004) conducted a detailed empirical study with variance ratios of 2.25 to 6.25 for very small sample sizes exploring Type I error accuracy in

conventional statistical tests, both population means and variance equality. A more

comprehensive resource, Coombs, Algina, and Oltman (1996), conducted a meta-analysis

of univariate and multivariate procedures noting similar variance ratios referenced in

other studies. The interest is only on the range of ratios as results of test statistic

accuracy in probabilistic inferences do not relate to an information based approach.

Turning to another pilot study, a series of variance ratios under fixed z^*, ϕ, and n_T

conditions evaluated performance of correct model selection and each of the three

hypotheses conclusions. The results in Table 11 indicate substantial differences most

noticeable in the mixture hypothesis.

Table 11.
Exploration of Variance Ratios for Normal Distributions on Model Level Performance

Population Conditions*			Empirical Results (%)**	
Variance Ratio	Population Model	Correct Model Selection	Correct Hypothesis Conclusion	
			Mixture	Variance
0.25	Model #5	86.6	86.8	99.8
0.33	Model #5	72.2	73.8	98.4
0.50	Model #5	32.0	40.6	91.4
1.00	Model #6	60.2	68.0	70.2
2.00	Model #5	96.2	96.2	98.2
3.00	Model #5	99.6	99.6	100
4.00	Model #5	100	100	100

* Note: All conditions were fixed with $z^* = 15$, $\phi = 0.20$, and $n_T = 200$.
** Note: Correct hypothesis conclusions regarding difference in population means were 100% for each experimental condition.

The results provide a snapshot for only a single set of conditions, yet are informative

enough to suggest variance ratios of {0.33, 0.50, 2.0, 3.0} be explored.

To keep the total number of experimental conditions at 580, only a subset of

conditions evaluated under population variance equality were considered. Z^* and ϕ

values remain unchanged, where n_T levels were reduced to

$$\{100, 200, 350\} \tag{101}$$

for variance ratios of 0.50 and 2.0. Using the same controlled parameter values enables comparison of all performance measures with the previous section. Unlike experimental conditions with variance equality, Model #1 is the correct specification with a zero mixing proportion and Model #5 otherwise.

For the more extreme variance ratio conditions, 0.33 and 3.0, an even smaller subset of conditions were evaluated. Z^* remains unchanged, with the ϕ levels reduced to

$$\{0, 0.10, 0.20, 0.35\} \tag{102}$$

and n_T levels reduced to

$$\{100, 200\} \tag{103}$$

Chapter 3 will synthesize the study results in both tabular and graphical fashion showcasing relationships between the controlled parameters, population conditions, and the 13 performance measures presented in section 2.3.

Chapter 3: Results

Chapter 2 developed the z^* composite parameter for systematic variation in a comprehensive study. Significant time was also spent presenting pilot study results of equal variance normal population conditions that demonstrating z^* as predictive in selected model level performance measures. Completion of the more comprehensive study including unequal variance conditions reaffirm z^* as instrumental and predictive across the entire set of performance measures of interest. This extends beyond select model level measures in the pilot study to include population and individual level measures. The remainder of the chapter focuses on characterizing these predictive relationships, which become additionally contingent upon the population variance ratio.

Some additional comments are necessary prior to the presentation of results. The empirical study was tremendously computationally intensive, requiring continuous employment of between 2 and 5 dedicated PCs and months to obtain the empirical results substantiating these findings. Summation across the 580 experimental conditions exceeded 250,000 trials, fitting roughly 1.75 million models where 80% were knowingly misspecified. To maximize the informative value from such an effort, some 13 performance measures were identified to more completely quantify the term *success*. Complicating the presentation for any of these performance measures is how to address results in the context of four important variables; z^*, ϕ, n_T, and σ^2_{rat}. Graphical displays are limited allowing variation of no more than two of these variables, requiring the remaining two to be assigned fixed values. Clearly, graphical results will change when the fixed variables are assigned different values, though care was taken to verify the general findings noted were consistent across the range of parameters values explored in

the study. Exacerbating this issue with the large number of performance measures identified, this analysis, while thorough, is far from exhaustive. Complete tabular results can be provided upon request to facilitate supplemental analysis or new analysis between performance measures within or across levels. The principal interest in this study, however, concerns successful hypotheses conclusions (treatment effect, mixture, and variance equality) assessed in conjunction with model selection. Complete tabular results of these conclusions are provided in Appendices 2-5, with more detailed analysis provided in this chapter. These findings also serve as the basis for a power analysis framework presented in the next chapter.

In evaluating each of the performance measures, z^* serves as the central parameter on which findings are characterized. Z^* exhibits asymptotic properties such that as

$$z^* \to \infty \tag{104}$$

nearly all hypotheses conclusions are 100% correct, estimates for parameters of interest become unbiased with decreasing variability, and correct individual group classification approaches 100%. Exceptions to this asymptotic result occur with conditions of equal population variance for the hypothesis of variance equality. The paths en route to asymptotic convergence across all performance measures as a function of z^* are non-linear, in some cases non-monotonic, and are conditioned on σ_{rat}^2 and ϕ, yet surprisingly invariant to n_T. Not invariant in the sense that n_T does not affect any of these performance measures, but n_T has no additional impact beyond it use in the calculation of z^*. From Equation (63), one also observes z^* is not independent of the σ_{rat}^2 and ϕ parameters. Unfortunately, large values of z^* are often not available in real world studies, where the characterization of these relationships takes on even greater

importance. Of course, a researcher can realize increased z^* values by a larger difference in population means, smaller variances in the populations, and larger sample sizes.

3.1 Model Level

Selection of the correct model specification is obviously the desired outcome, which in turn results in correct treatment effect, mixture, and variance equality hypotheses conclusions. However, this measure alone is insufficient as Table 5 illustrated a number of incorrect model selections which still provide correct conclusions to one or more hypotheses. Such possible outcomes necessitate separate analysis for each hypothesis. An overarching model level measure was the successful convergence of the *SQP* search algorithm supporting the ML process for the three models requiring its use per trial. The results were quite surprising; for each model, for every trial, for each empirical condition, irrespective of the extent of misspecification, there was a 100% convergence to a viable solution also satisfying specific model fit requirements. The implication is all six models provided computational *AIC* values enabling a complete set for a min *AIC* selection. Whether a testament to algorithmic efficiency or the flexibility of normal distributions in finite mixtures, this result dispels any notion that successful convergence is a type of evidence in support of correct model specification. More informative convergence information, such as tracking the number of attempts per trial and documenting the reason for any convergence failures, was not retained.

3.1.1. Correct Model Selection

Figure 21 illustrates the asymptotic properties of z^* on correct model selection,

with distinctive rates conditioned on the variance ratio. Figure 9 from the pilot study

foreshadowed these findings where additionally results were not symmetrical as the

variance ratio deviated from 1.0. Larger variance ratios, 2.00 and 3.00, consistently

outperformed the others, while variance ratios of 1.00 and 0.50 comparatively performed

the poorest.

Figure 21.
Graph: Correct Model Selection at $\phi = 0.20$, $n_T = 200$ over z^* by σ^2_{rat}
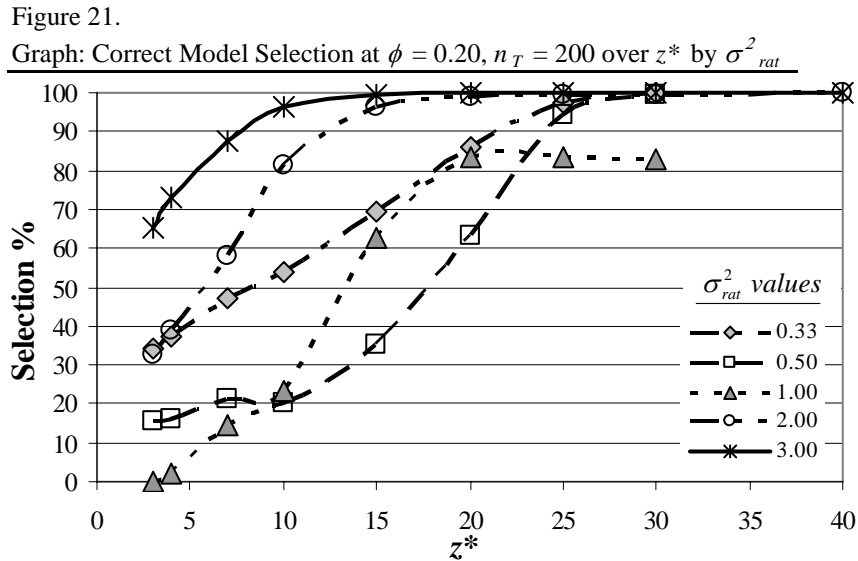
The exception to the asymptotic convergence of z^* in correct model selection occurs at

population conditions of variance equality. In such cases, this process asymptotes around

83%, where the limitation, as shown later, was in selecting models of equal variance. Of

note, this was the only exception in the asymptotic properties of z^* across all performance

measures analyzed in this study. The findings were consistent at other ϕ and n_T

specifications.

Figure 22 evaluates model selection with a variation of ϕ with fixed z^* and n_T

values. Improved model selection solely as result of an increase in ϕ does not

consistently hold, as in the case of a σ^2_{rat} of 2.0 where selection decreases slightly from

its peak at $\phi = 0.20$. Figure 11 from the pilot study similarly demonstrated a slight

decrease in correct model selection for larger ϕ values with a different set of fixed

conditions. As possible explanation for this phenomenon, recall Figure19 when equal

sample sizes are used, the difference between z^* and z^*_{max} increases dramatically with

larger ϕ values. Explained from a different perspective, the effective sample size ratio,

Equation (91), deviated greatly from 1.0. Variance ratios of 0.50 and 1.00 comparatively

performed worst among variance ratios evaluated. The results were consistent when

other fixed z^* values were used.

Figure 22.
Graph: Correct Model Selection at $z^* = 10$, $\phi > 0.00$, $n_T = 100$ over ϕ by σ^2_{rat}
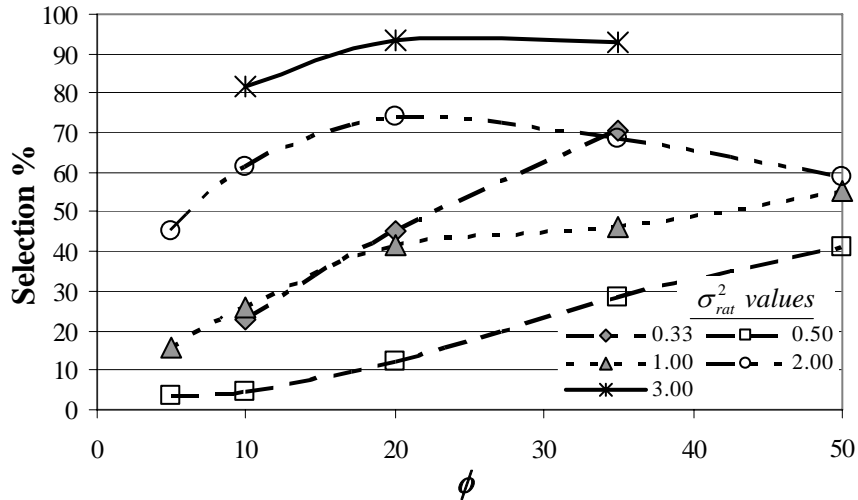
Figure 23 considers correct model selection for the poorest performing variance

ratio, 0.50, considering the effects of sample size at a fixed ϕ. Results were mixed. At

lower z^* values, an increase in sample size is more beneficial to correct model selection,

while the opposite occurs for larger z^* values. It is impossible to distinguish a separate

impact for sample size, using n_T as the multi-sample measure, when results change over

the range of z^*.

Figure 23.

Graph: Correct Model Selection at $\sigma^2_{rat} = 0.50$, $\phi = 0.10$, over $z*$ by n_T
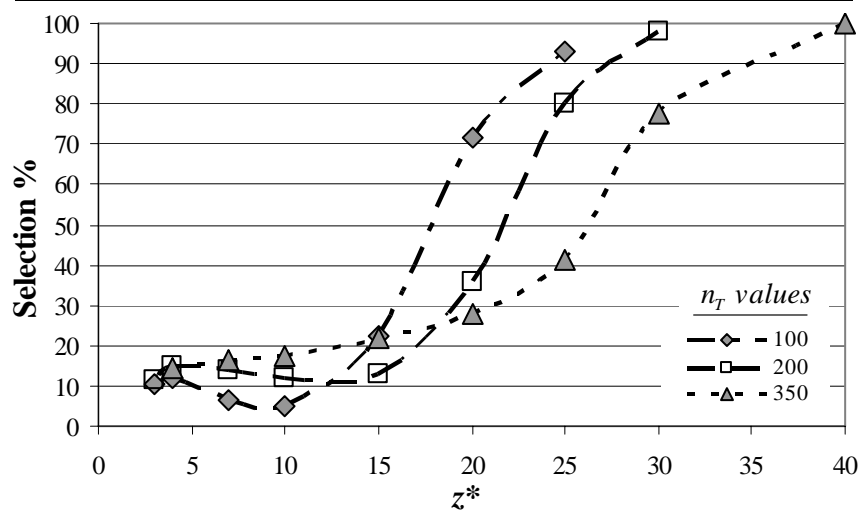
3.1.2 Treatment Effect Hypothesis Conclusion

The first hypothesis conclusion to be evaluated is whether model selection is accompanied with a treatment effect result. All empirical conditions had a population treatment effect, so instances of a false selection could not be evaluated. No treatment effect corresponds to equality in population means where the d_u, d^*, and z^* values are 0. Tabular results are provided in Appendix 2, where any z^* value exceeding 7, regardless of σ^2_{rat}, ϕ, and n_T value had a correct treatment effect hypothesis conclusion rate of 100%. Figure 24 is a 3D graph varying variance ratio and z^*, only showing the smaller z^* results. Among the variance ratios evaluated, a σ^2_{rat} of 1.0 had the worst performance. The results appear fairly symmetrical as σ^2_{rat} of 0.50 and 2.0 has similar results and the best performance occurred at the largest variance ratio deviations from 1.0.

Figure 24.

Graph: Correct Treatment Effect Hypothesis Conclusion by $z*$ and σ^2_{rat}

$\quad\quad$ ($\phi = 0.10$ and $n_T = 100$)
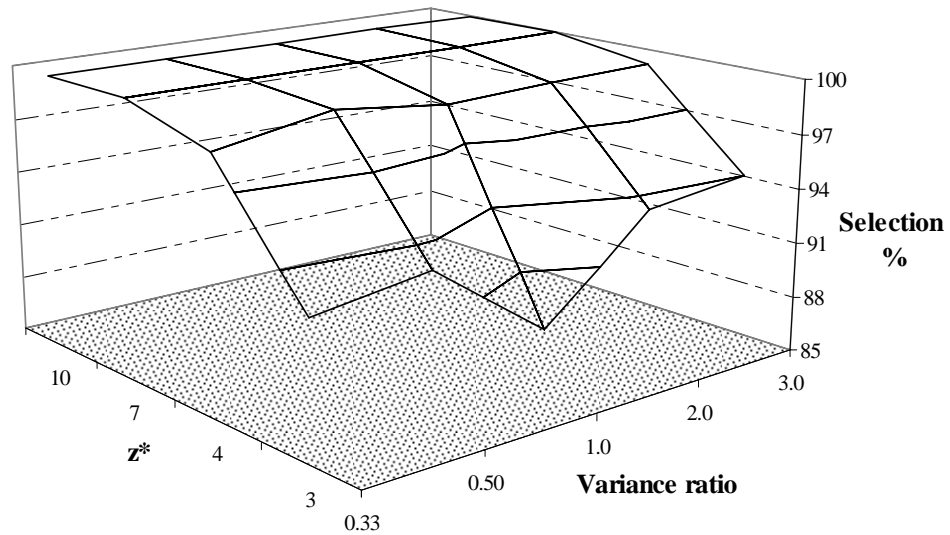
Figure 25 varies z^* and now ϕ for the worst performing variance ratio with respect

to a correct treatment effect hypothesis conclusion, 1.0. Across the range of ϕ values,

selection goes to 100% as z^* increases.

Figure 25.

$\quad$ Graph: Correct Treatment Effect Hypothesis Conclusion by $z*$ and ϕ

$\quad\quad$ ($\sigma^2_{rat} = 1.0$ and $n_T = 200$)

80
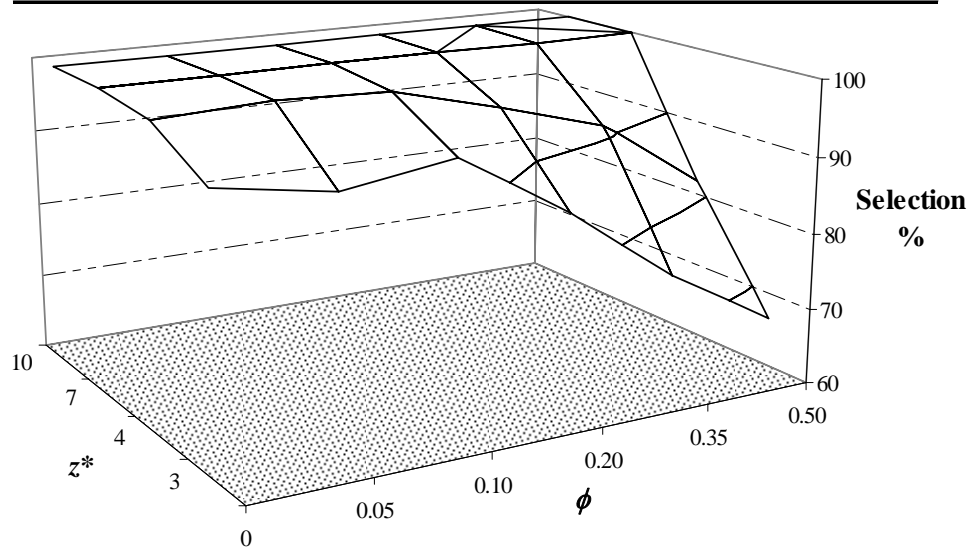
The correct treatment effect hypothesis conclusion decreases with larger ϕ values, which may similarly be attributed to increasingly larger deviations from 1.0 in the effective sample size ratio. With the same result for all experimental conditions where z^* exceeds 7, there was not sufficient remaining data to explore differences in this hypothesis conclusion as a result of changes in n_T for fixed z^* and ϕ values.

3.1.3 Mixture Hypothesis Conclusion

The most important empirical evidence in any finite mixture model concerns the population mixing proportion estimate, $\hat{\phi}$, commensurate with a min *AIC* model selection. Quantifying this performance measure comes from two directions: the false selection of a mixture model when no treatment nonresponse was present and correct detection of a mixture via model selection. Complete tabular results of false mixture model selection rates are provided in Appendix 3 while Appendix 4 provides the complete tabular results of a correct mixture hypothesis conclusion when treatment nonresponse was present.

False selection rates examine experimental conditions without treatment nonresponse, hence a ϕ of 0.0, where results are graphically presented in Figure 26 simultaneously varying z^* and σ^2_{rat} for a fixed n_T.

Figure 26.

Graph: False Mixture Hypothesis Conclusion by $z*$ and σ^2_{rat} ($\phi = 0.0$ and $n_T = 200$)
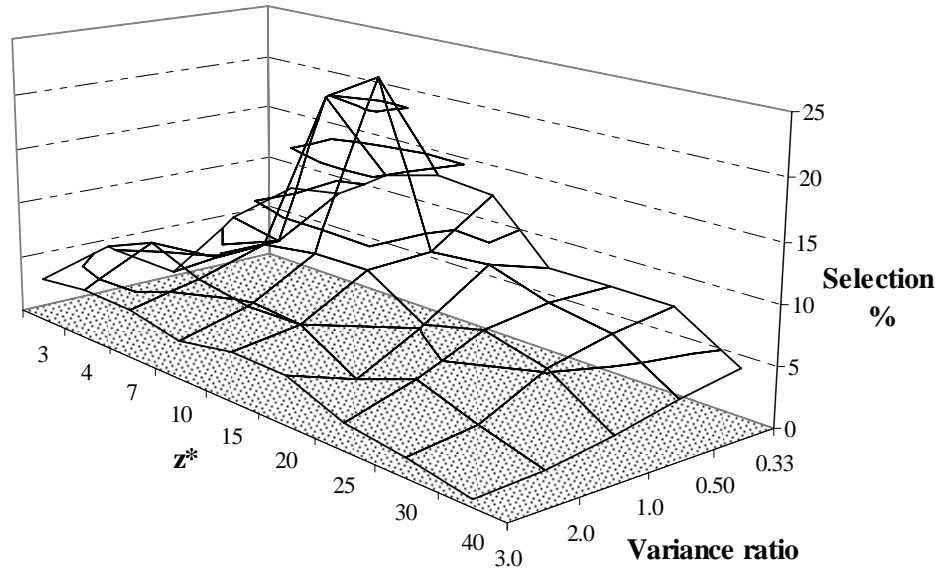
An overall assessment is higher variance ratios performed better where even the σ^2_{rat} of 1.0 outperformed ratios below 1.0. False selection rates are noticeably higher in population conditions having a σ^2_{rat} of 0.50, particularly in the z^* range of 7-15. Across all variance ratios, false selection rates were lowest at the extremes of z^* in the study, both small and large, making the relationship nonlinear. Lower false selection rates at small z^* values are attributed to the decreased selection of models containing a treatment effect, which is a prerequisite for evidence of treatment nonresponse. Lower false selection rates at higher z^* values is attributed to the asymptotic properties of z^*, a commonly observed attribute as different measures are subsequently evaluated.

Analysis of the correct mixture hypothesis conclusion with treatment nonresponse was more involved due to the additional systematic variation of ϕ. Figure 27 varies the same parameters as Figure 26 where the experimental conditions now have a fixed ϕ of 0.20.

Figure 27.

Graph: Correct Mixture Hypothesis Conclusion by z^* and σ^2_{rat} ($\phi = 0.20$ and $n_T = 100$)
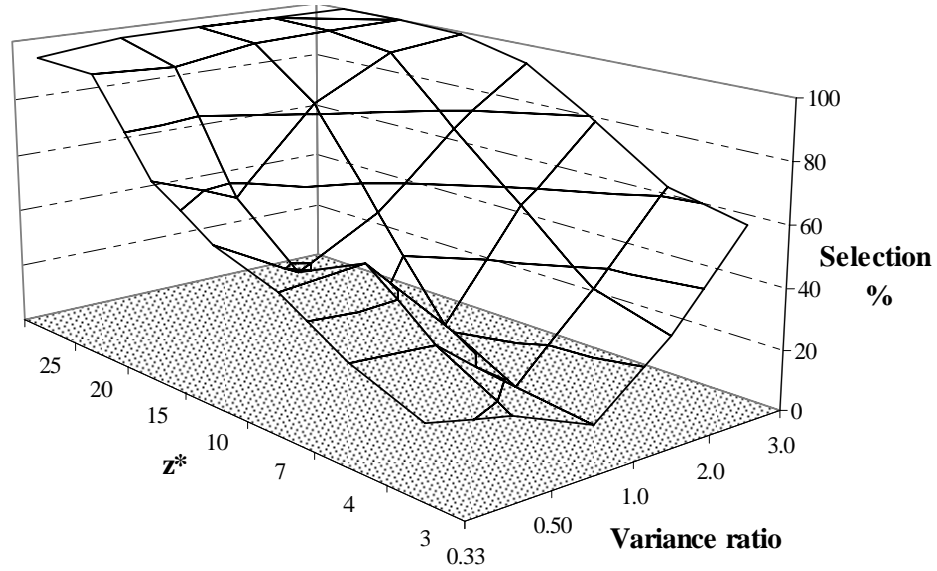
Higher variance ratios had the highest correct mixture hypothesis conclusion rates, a consistent performance finding across the performance measures. Across all σ^2_{rat}, correct hypothesis conclusion rates move to 100% as z^* increases demonstrating asymptotic properties for a different measure. Among the variance ratios, σ^2_{rat} of 1.0 performed poorest. Also, the changes in correct hypothesis conclusions were not monotonic for variance ratios below 1.0. This result is more clearly seen in a 2D graph, Figure 28, which used different fixed ϕ and n_T values from Figure 27.

Figure 28.

Graph: Correct Mixture Hypothesis Conclusion at $\phi = 0.10$, $n_T = 100$
over $z*$ by σ^2_{rat}
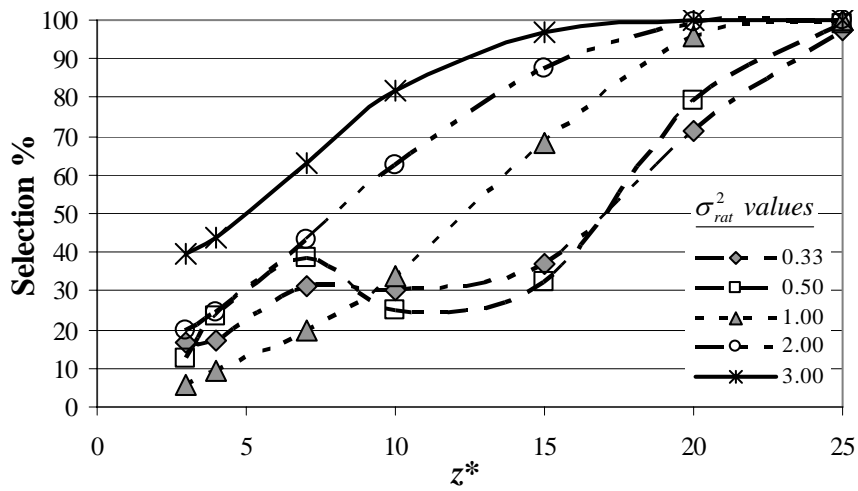
The findings are consistent with the previous figure illustrating higher selection rates for population variances exceeding 1.0 and the asymptotic properties of z^* on the correct treatment effect hypothesis conclusion. For these experimental conditions, however, a σ^2_{rat} of 1.0 was not the poorest across the entire range of z^*, replaced by the lowest variance ratios as z^* increased.

To evaluate the impact of ϕ separate from z^*, ϕ was varied for two fixed z^* and n_T values presented side by side in Figure 29. In each instance, the larger z^* values resulted in higher selection rates when comparatively assessing like variance ratios. As with the correct model selection measure, selection rates for a mixture hypothesis tend to improve with larger ϕ values, but that was not a uniform result with exceptions in the largest ϕ values evaluated.

Figure 29.
Graphs: Correct Mixture Hypothesis Conclusions at $z* = 7$ and 15 over ϕ by σ^2_{rat} ($n_T = 200$)
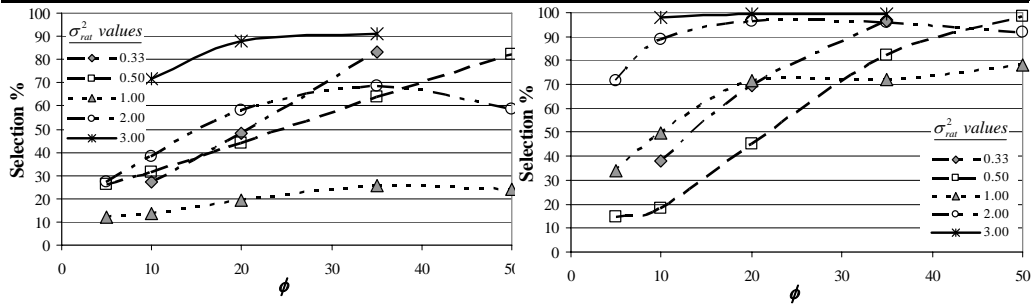
The variation of ϕ can also be represented in a 3D graph also varying z^* for a single σ^2_{rat} value. For this particular set of conditions, larger ϕ values resulted in higher correct mixture hypothesis conclusions compared to conditions of less treatment nonresponse.

Figure 30.
Graph: Correct Mixture Hypothesis Conclusion by $z*$ and ϕ ($\sigma^2_{rat} = 0.50$ and $n_T = 200$)
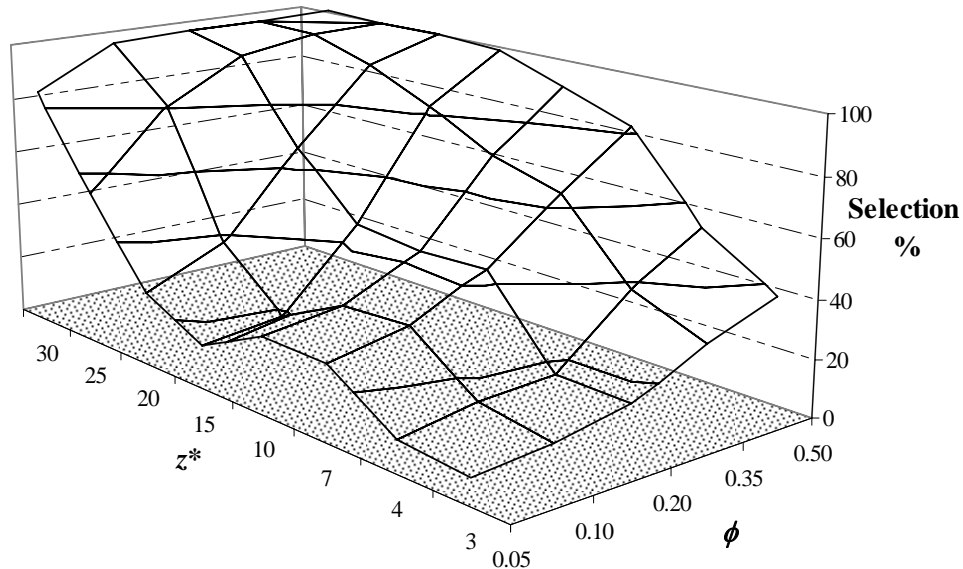
Another notable result is the lack of impact, or invariance to n_T beyond its use in the z^* calculation on the correct mixture hypothesis conclusion. This is clearly illustrated in Figure 31 using a σ^2_{rat} of 2.0 and a fixed ϕ of 0.05. Other variance ratios and fixed ϕ values were evaluated to validate this finding. The invariance of n_T may hold as well for

the treatment effect hypothesis, but there was not enough information from the study to either support of refute that claim.

Figure 31.

Graph: Correct Mixture Hypothesis Conclusion at $\sigma^2_{rat} = 2.00$, $\phi = 0.05$ over $z*$ by n_T
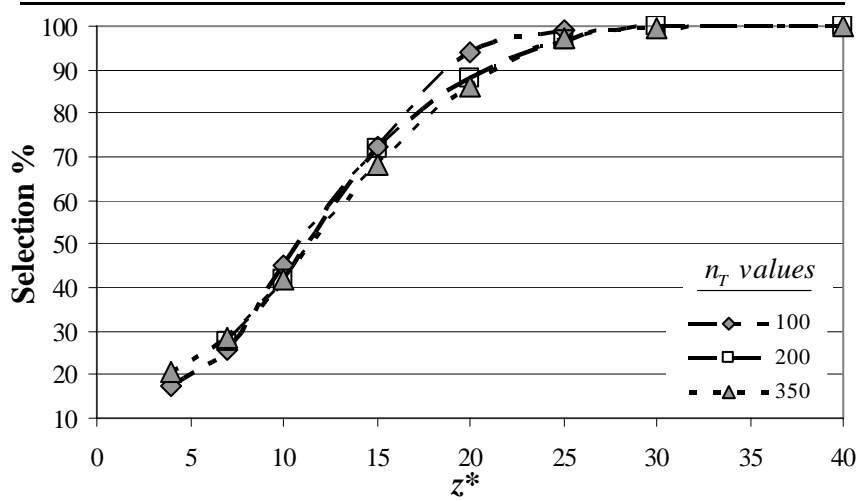
While correct model selection has undeniable value, hypotheses conclusions are reported and, therefore, a more important result. From Table 5, correct model selection rates serve as the lower bound of the correct conclusion rates for each of the three hypotheses posited. To explore this relationship with the mixture hypothesis, Figure 32 provides results for a set of nonzero ϕ conditions by variance ratio, where the dashed diagonal line graphically illustrates the lower bound property.

Figure 32.
Graph: Correct Mixture Hypothesis Conclusion at $0 < \phi <= 0.20$, $n_T = 200$
over Correct Model Selection by σ^2_{rat}
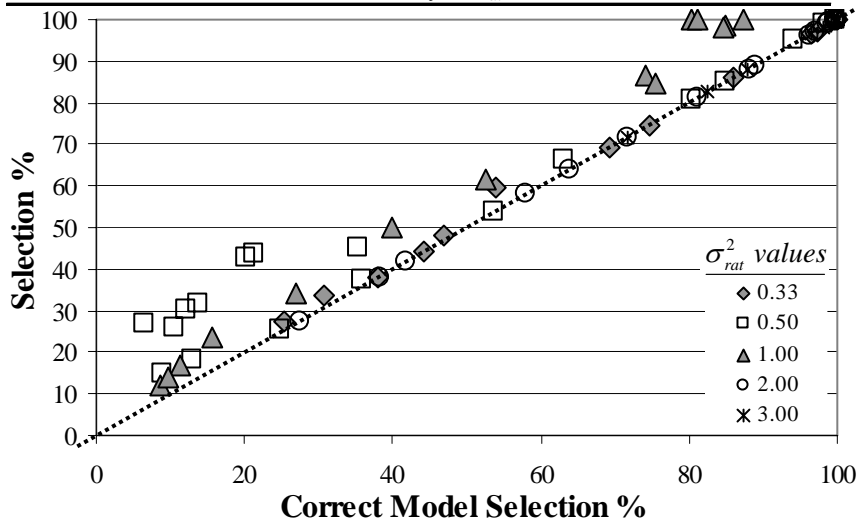
In every case, variance ratios exceeding 1.0 had exact correct model and correct mixture

hypothesis conclusion selection rates. That was not the case for the other variance ratio

values where greatest discrepancies occurred at σ^2_{rat} of 0.50 and 1.0. For the σ^2_{rat} of

0.50, the larger discrepancy occurs at low model selection rates and diminishes as the

correct model selection improves. The opposite was true for the σ^2_{rat} of 1.0, where the

discrepancy increases with higher correct model selection rates.

3.1.4 Variance Hypothesis Conclusion

Tabular results for all experimental conditions are provided for this hypothesis in

Appendix 5. Using the same process as the other hypotheses, Figure 33 varied both z^*

and the σ^2_{rat} at fixed ϕ and n_T conditions.

Figure 33.

Graph: Correct Variance Hypothesis Conclusion by $z*$ and σ^2_{rat} ($\phi = 0.20$ and $n_T = 200$)
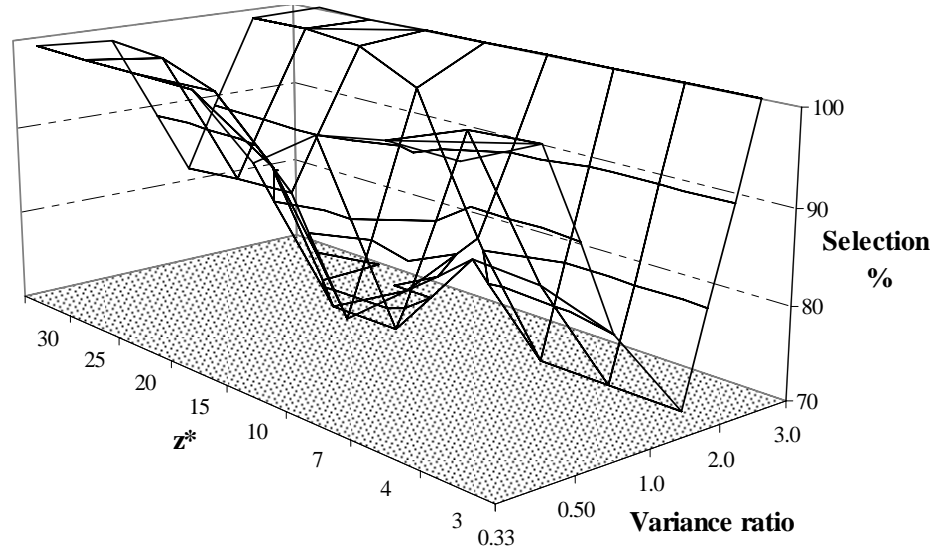
The farther the population variance ratio deviates from 1.0, the higher the rate of correct

hypothesis conclusion. For σ^2_{rat} of 3.0 and 0.33, correct hypothesis selection rates are

consistently at or near 100% over the range of $z*$ values evaluated. For σ^2_{rat} of 2.0 and

0.50, correct hypothesis selection rates consistently increased toward 100% as $z*$

increased. However, for σ^2_{rat} of 1.0, correct hypothesis selection rates never exceeded

83% regardless of the increase in $z*$ value. These results indicate it is more difficult to

provide empirical evidence through model selection supporting variance equality than

evidence supporting variance inequality.

To better demonstrate the asymptotic properties of $z*$ for this hypothesis with

variance ratios other than 1.0, Figure 34 varies both $z*$ and ϕ for a σ^2_{rat} of 2.0. For any ϕ,

the selection rate improves as $z*$ increases. In comparison with the mixture hypothesis,

the opposite result of lower selection rates occurs as ϕ increases.

Figure 34.

Graph: Correct Variance Hypothesis Conclusion by z^* and ϕ ($\sigma^2_{rat} = 2.0$ and $n_T = 100$)
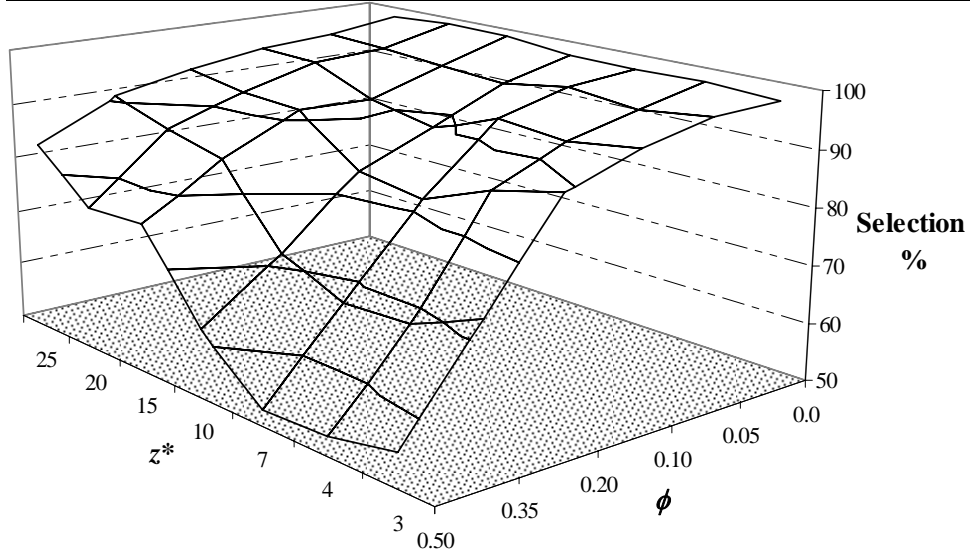
Figure 35 serves two purposes: to evaluate the effects of z^* for a σ^2_{rat} of 1.0 and

demonstrate its invariance to n_T beyond its use in the calculation of z^*. As shown, in z^*

regions of 7-20, the correct variance hypothesis conclusion rates drop below 70%, but at

the extremes of the z^* ranges evaluated, the rates climb to 83%. Further, these results

were consistent irrespective of the n_T values, which ranged from 50 to 350. Examination

of other variance ratios and ϕ conditions similarly showed no change in hypothesis

conclusion rates for differing n_T values.

Figure 35.

Graph: Correct Variance Hypothesis Conclusion by z^* and n_T ($\sigma^2_{rat} = 1.0$ and $\phi = 0.35$)
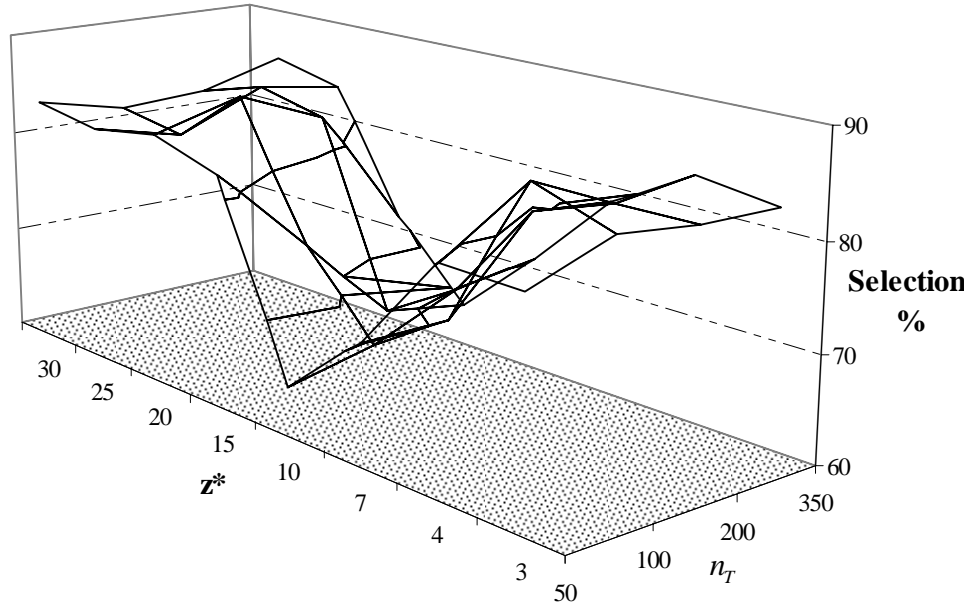
3.2 Population Level

3.2.1 Interpretability of Parameter Estimation Characteristics

Estimation of model parameters within a maximum likelihood framework is well established as an effective and reliable method. Perhaps the greatest feature of this method, aside from being generally scale invariant, is its robustness in yielding quality parameter estimates (Kaplan, 2000). This finding has been supported in a number of studies including Anderson (1988) and Olsson (1999). The hallmarks of quality in a parameter estimation technique are accuracy and consistency. The statistical analog to the accuracy attribute is *bias*, shown in Equation (69). Statistical analogs commonly associated to the consistency attribute are either variability or *MSE*, shown in Equations (72) and (71) respectively. In these same references, and more generally, a strong

assumption accompanying the use of maximum likelihood and inferential testing on

parameter estimates is correct model specification. The min *AIC* strategy advocated is a

data driven model selection using information criteria.

In a series of articles, Leeb (2005, 2006) clearly articulated the difficulties and

even impossibilities in determining parameter estimates' distributional properties to

enable inference in conjunction with a model selection process. While Leeb's work did

not include finite mixture models, his work is especially informative in the context of this

problem. Unlike his work, this interest is not making probabilistic inferences on the

model parameters or parameter composite estimates, $\hat{\phi}, \hat{d}_u$, or $\hat{d}^*$, but at a methodological

level to defend this entire process as consistent and unbiased. Even at this more basic

level, each experimental condition evaluated resulted in a different amalgamation of

correct and different incorrectly specified models from a set of trials which form the *bias*

and *MSE* statistics. Issues surrounding model misspecification and maximum likelihood

have been discussed in the literature for at least 35 years, with more recent contributions

particular to the area of finite mixture models (White, 1982; Gray, 1994). These articles

focus on the strong assumption failure in correct model specification where the same

wrong model is repeatedly fit. A min *AIC* strategy forgoes that ML assumption, relying

instead on information criterion assessing overall model fit.

With different compositions of models in each experimental condition, an overall

assessment of this process in regards to the traditional metrics of *bias* and *MSE* similarly

reaches Leeb's conclusion as 'an impossibility'. If the issue of varied model composition

per experimental condition was not enough, even the use of *bias* and *MSE* as viable

measures is in question. It is reassuring and natural to envision these terms graphically as

nicely unimodal, symmetric shapes centered or only slightly off center the true

population value. Without these properties, these measures have little to no interpretative

value. While varied model composition in each experimental condition reduces the

likelihood of attaining these properties, there are other factors which also degrade the

interpretability of these parameter estimation measures.

a. Not all model parameters are freely estimated from the ML process. Among

the competing models, different parameter constraints exist between samples

for variance and/or mean equality.

b. The mixing proportion parameter estimate is bounded by definition and further

constrained in model specification where

$$0 < \hat{\phi} < 1 \tag{105}$$

c. The reported treatment effects estimates of interest, $\hat{d}_u$ and $\hat{d}^*$, are not directly

estimated, but are composites of separate model parameter estimates that

computationally involve an absolute value transformation. These composites

are bounded in construct such that

$$d_u, d^* \geq 0 \tag{106}$$

3.2.2 Experimental Condition Examples Illustrating Summary Parameter

Estimation Measures

To exemplify these difficulties, four empirical conditions from among the 580

were chosen varying σ^2_{rat} and z^* values with a wide range of correct model selection and

hypotheses conclusion rates. These experimental conditions and their corresponding

model level performance results are provided in Table 12.

Table 12. Experimental Conditions and Model Level Results used in Parameter Estimate Histogram Assessment

Experimental Conditions	Population Values						Min AIC Model Results[*] (%)		
	σ^2_{rat}	ϕ	z^*	n_T	d^*	d_u	Correct Model	Correct Mix Hyp	Correct Var Hyp
Condition #1	1.00	0.20	10	50	2.04	10	54.8	65.4	72.2
Condition #2	2.00	0.20	10	200	0.95	10	81.2	81.2	98.6
Condition #3	3.00	0.20	15	100	1.96	10	99.0	99.0	99.4
Condition #4	0.50	0.20	15	350	1.24	10	42.0	51.4	90.6

* Note: All experimental conditions resulted in a 100% correct hypothesis conclusion regarding difference in population means.

Instead of focusing on each individual parameter estimate, analysis is limited on recapturing the population mixing proportion, ϕ, and the unstandardized and standardized treatment effect representations, d_u and d^*. Of the three, ϕ could be characterized as the simplest because it is not a composite representation of various estimates. Figure 36 provides a histogram of the selected conditions $\hat{\phi}$ values where 0.20 was the true population value.

Figure 36.
Graph: Histogram of ϕ_{est} for Selected Experimental Conditions
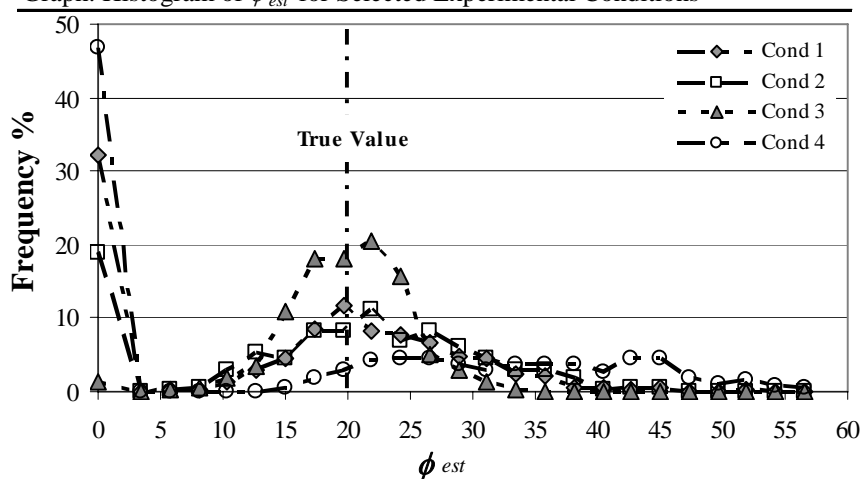
Failure to arrive at the correct mixture hypothesis conclusion, not the more restrictive correct model selection, creates a bimodal shape. Not only do bimodal shapes

confound the interpretation of *bias*, but increase the estimate's variability captured in the

MSE term. Such conditions also make any type of probabilistic inference on the estimate

untenable, where only condition #3 presents a unimodal shape. Yet, in this particular

case, the results are not symmetric, having a negative skew. While not as problematic as

bimodal representations, the accuracy of any probabilistic inference becomes an issue.

Either bimodality or skewness creates a difference between central tendency and the

average estimate.

Turning to the treatment effect representations, there was no issue of bimodality,

resulting from the 100% correct treatment effect hypothesis conclusions for each

condition. Examining the standardized treatment effect in Figure 37, even with error free

hypothesis conclusions, declaring the histograms of these estimates as neatly unimodal is

an overly generous characterization. Not only are probabilistic inferences inappropriate,

interpreting *bias* is challenging where Equation (77) uses an estimate average, but the real

interest is in the central tendency. Because results can not be attributed to incorrect

hypothesis conclusions, the most likely cause is its composite representation of many

parameter estimates in a complex formula, Equation (57).

Figure 37.
Graph: Histogram of $d*_{est}$ for Selected Experimental Conditions

94
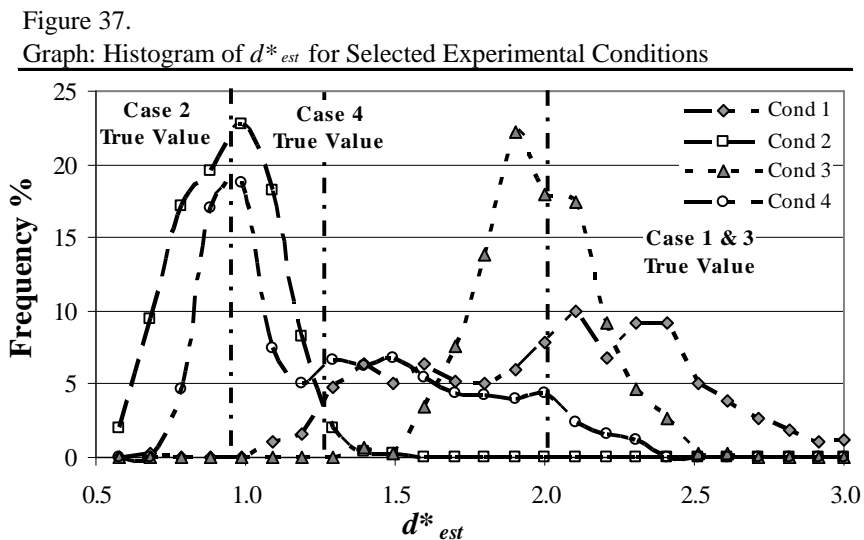
A simpler composite representation is the unstandardized treatment effect, which only involves the population mean estimates. Benefiting as well from the 100% correct hypothesis conclusion, compared to Figure 37, Figure 38 has more consistent results across the selected conditions centered on the true value. The results are far from ideal, where only condition #3 demonstrates sufficient unimodal and symmetric properties suitable to enable probabilistic inference. Further, the process and model selection compositions are such that making generalizations across all experimental conditions is ill-advised.

Figure 38.
Graph: Histogram of $d_{u\,est}$ for Selected Experimental Conditions
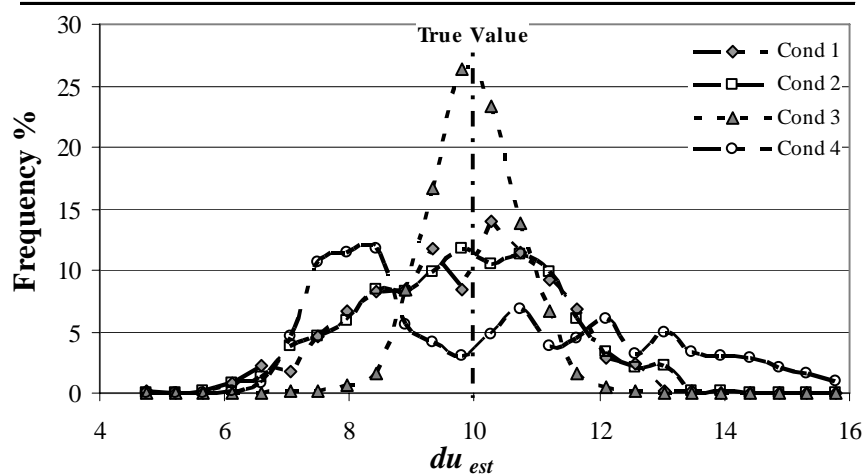
Unfortunately, despite a strong set of arguments with supporting empirical evidence critiquing the value of *bias* and *MSE* as population levels measures of performance under this methodology, no other alternative is available. Performing no analysis and thereby ignoring the quality of the population level estimates is undoubtedly the worst course of action. There is no value in obtaining correct hypotheses conclusions only to report wildly inaccurate results. So analysis must proceed, but cautiously, using

bias and *MSE* qualifying the recapturing of these selected population parameters. To appropriately describe the goals of this analysis:

a. Given the nebulous value of the *bias* and *MSE* terms used, their use in assessing this methodology is more qualification than quantification.

b. Probabilistic inferences for any of these parameter estimates are not recommended. Fortunately, because this methodology operates at the model level through a comparative evaluation of many models, such inferences are not necessary.

c. Z^* will be used as the independent variable in evaluating population level measures, having demonstrated informative relationships with asymptotic properties and invariance to n_T for the three hypotheses posited.

3.2.3 Population Mixing Proportion Estimate

Figure 39 shows the $\hat{\phi}$ bias over z^* for each σ^2_{rat} in the empirical study. Similar to the model level measures, σ^2_{rat} values above 1.0 performed better. Biases are mostly negative, particularly at lower values of z^*, due to an incorrect mixture hypothesis conclusion. With an incorrect hypothesis conclusion, $\hat{\phi}$ is 0, which perpetuates a bimodal histogram shape in the set of trials, lowering the average estimate value used in determining the *bias* value in this figure.

Figure 39.

Graph: ϕ Estimate Bias at $\phi = 0.20$, $n_T = 100$ over $z*$ by σ^2_{rat}
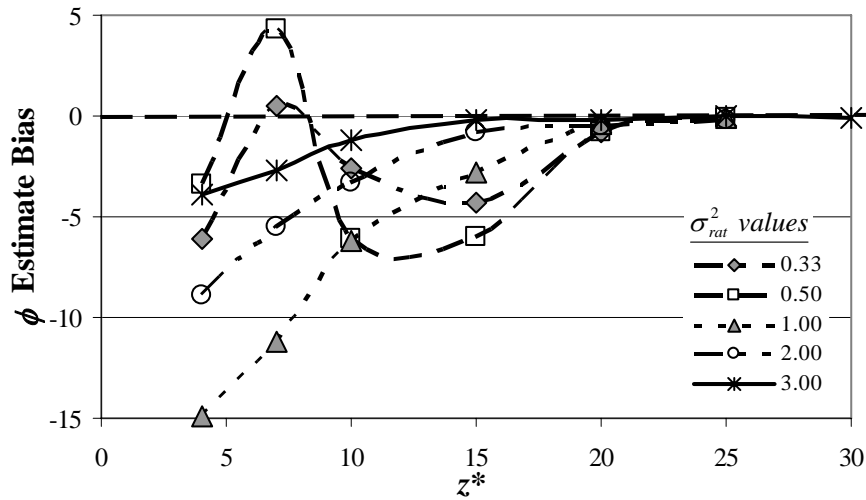
Demonstrated in the last section, higher z^* values result in a higher correct mixture hypothesis conclusion, lessening or removing the bimodal condition. Increasingly, the resulting shape of the estimate set will be unimodal, as shown in Figure 36, though are likely to be skewed and kurtotic to some extent. Under these conditions, there is interpretative value in the *bias* term where a value of 0 is most preferred. That the *bias* values differ conditioned on σ^2_{rat} as a function of z^* is interesting, but most important is the *bias* goes to 0 for each σ^2_{rat} value as z^* increases. These results were consistent under examination of other fixed ϕ and n_T conditions.

To examine the effects of varying sample size on the $\hat{\phi}$ bias, the correct mixture hypothesis conclusion was used as the independent variable with population conditions whose σ^2_{rat} was 1.0. Figure 40 shows the invariance of results to sample size, this time in respect to a population level measure where the strong relation between z^* and the mixture hypothesis conclusion has been established. Evaluation of other σ^2_{rat} and ϕ conditions yielded the same result.

97

Figure 40.

Graph: ϕ Estimate Bias at $\sigma^2_{rat} = 1.0$, $\phi = 0.20$ over Correct Mixture

Hypothesis Conclusion % by n_T
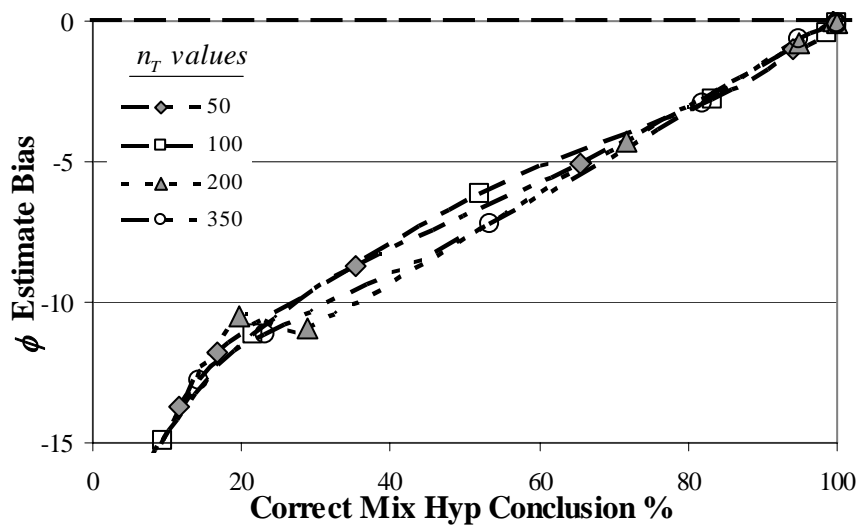
The measure of consistency used for $\hat{\phi}$ is *MSE*, combining both variability and

bias into a summary measure. The results for this measure, shown in figure 41 are

consistent with many of the previous measures evaluated: different relationships over z^*

as a function of σ^2_{rat} and comparatively better performance for σ^2_{rat} values exceeding 1.0.

Like the *bias* measure, the *MSE* of $\hat{\phi}$ decreases to 0 as z^* increases.

Figure 41.

Graph: *MSE* (ϕ Estimate) at $\phi = 0.10$, $n_T = 100$ over z^* by σ^2_{rat}
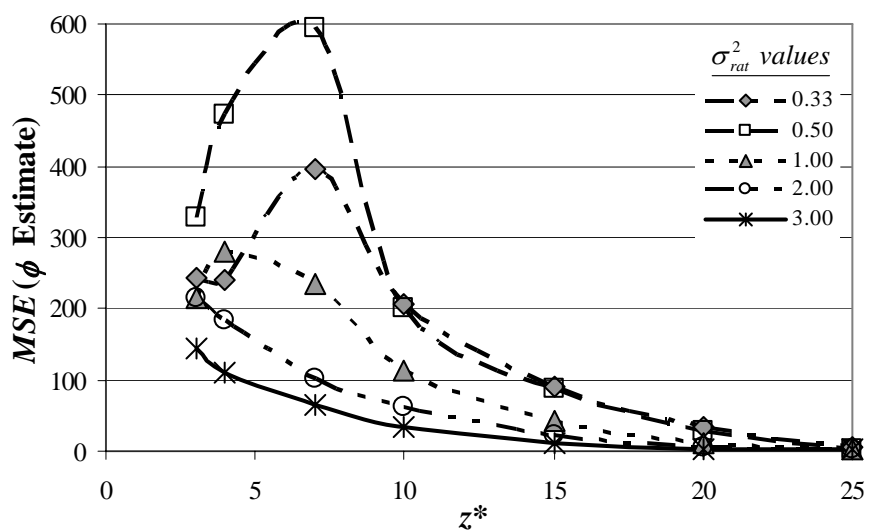
3.2.4 Population Treatment Effect Estimates

Qualifying the recapturing of true population treatment effect involves analysis of

two composite representation estimates, $\hat{d}_u$ and $\hat{d}^*$. While these estimates are not

afflicted with bimodality as a function of incorrect hypothesis conclusions, generalizing

results is confounded by their formulaic expression of many estimates. As a cursory

exploration, the scatterplot in Figure 42 was constructed involving these estimates' *bias*,

$\hat{\phi}$ *bias*, correct model selection, and the correct mixture hypothesis conclusion. Due to

the preponderance of experimental conditions with 100% correct treatment effect

hypothesis conclusions, this measure was uninformative in indicating any relationships

and was omitted from the scatterplot.

Figure 42.
Graph: Matrix Scatterplot of Selected Parameter Estimate Biases (all conditions at $n_T = 200$)
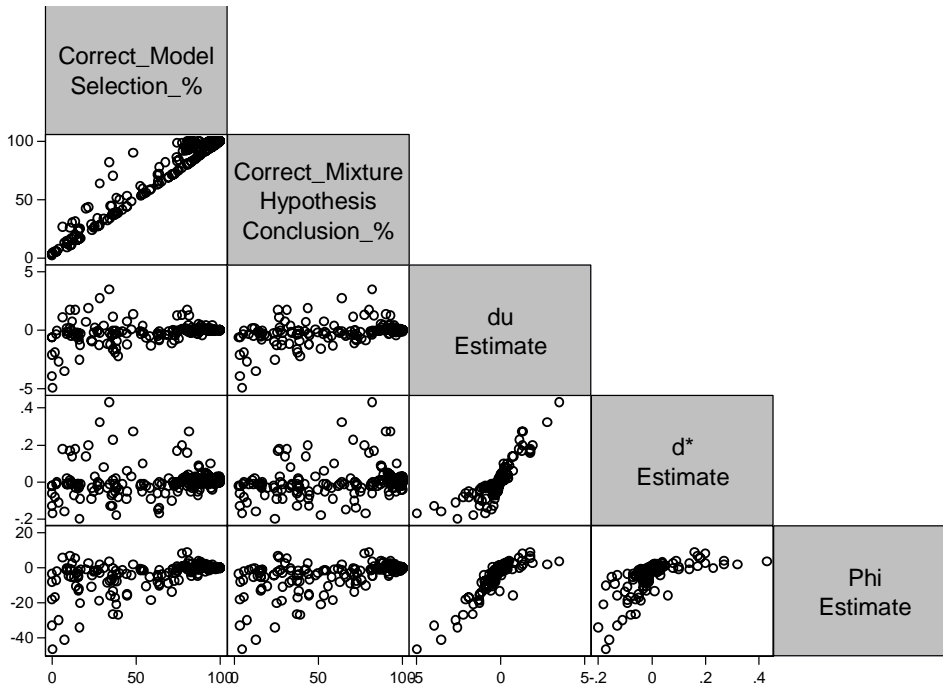
The uppermost left plot is a similar representation of Figure 32 where the σ^2_{rat}

values are not separately distinguished. Strong $\hat{d}_u$ and $\hat{d}^*$ correlations and similar

relationships to the other measure was expected, and while confirmed in this plot, were

hard to qualify with inclusion of additional parameter estimates to calculate $\hat{d}^*$. Where

previous analysis pinpointed examination of an estimate's variability for a single

experimental condition over a set of trials, these plots provide a broader view

characterizing variability in their estimate biases across many experimental conditions.

There appears to be no definitive relationship between correct model selection and either

treatment effect estimate *bias*, though the overall cone-like shapes do indicate a

decreased *bias* variability as model selection improves. Surprisingly, there is a stronger

relationship between $\hat{d}_u$ and $\hat{\phi}$ biases than between the two treatment effect

representations.

Returning to the variation of z^* with fixed ϕ and n_T values, Figure 43 shows $\hat{d}^*$

bias approaches 0 as z^* increases. Figure 44 illustrates similar findings using d_u estimate

bias as the treatment effect measure with a different set of ϕ and n_T conditions.

Figure 43.
Graph: *d** Estimate Bias at $\phi = 0.20$, $n_T = 100$ over z^* by σ^2_{rat}

100
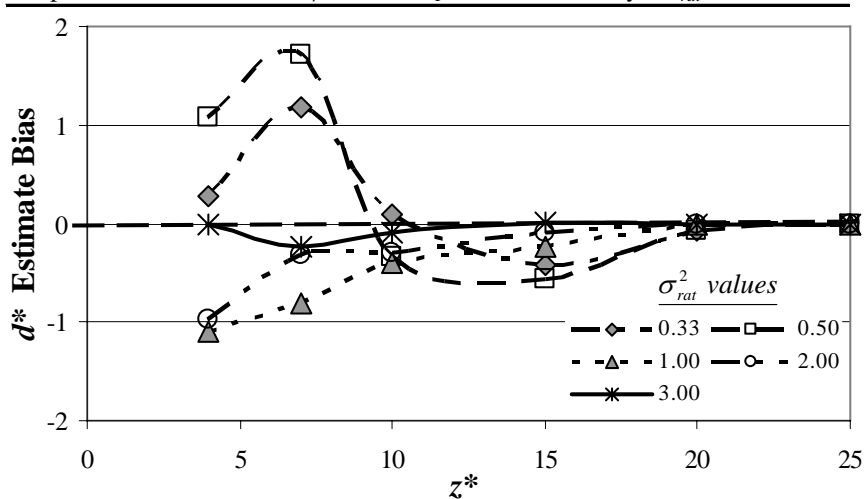
Figure 44.

Graph: d_u Estimate Bias at $\phi = 0.35$, $n_T = 200$ over $z*$ by σ^2_{rat}
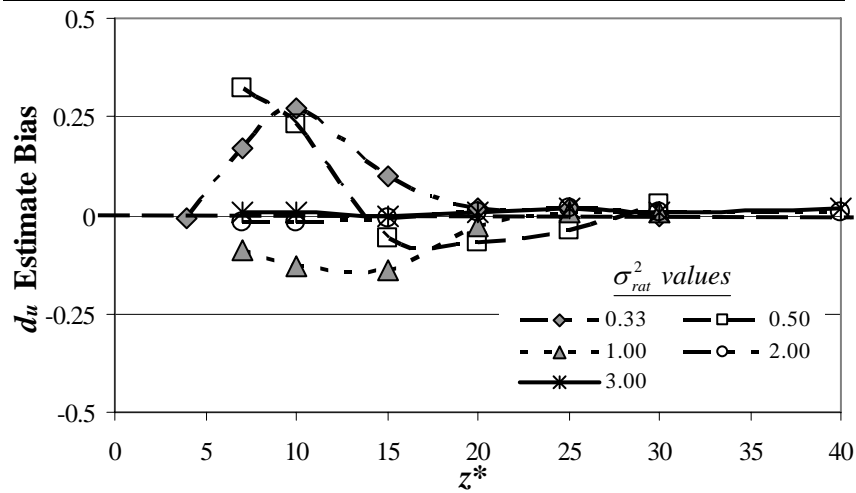
Reaffirmed with other ϕ and n_T conditions for both treatment effect representations, σ^2_{rat} values exceeding 1.0 performed best on these performance measures. An interesting observation while *bias* converges to 0, the treatment effect tends to overestimation for σ^2_{rat} conditions below 1.0 while tending to underestimation for the remaining σ^2_{rat} conditions.

Focusing of variability across experimental conditions, the *MSE* of the unstandardized treatment effect estimated was presented in Figure 45. Again, the asymptotic properties of z^* are displayed where the *MSE* of this estimate approaches 0 as z^* increases. Similar results occurred with consideration of the $\hat{d}^*$ *MSE*.

Figure 45.

Graph: $MSE\,(d_u$ Estimate) at $\phi = 0.10$, $n_T = 100$ over z^* by σ^2_{rat}
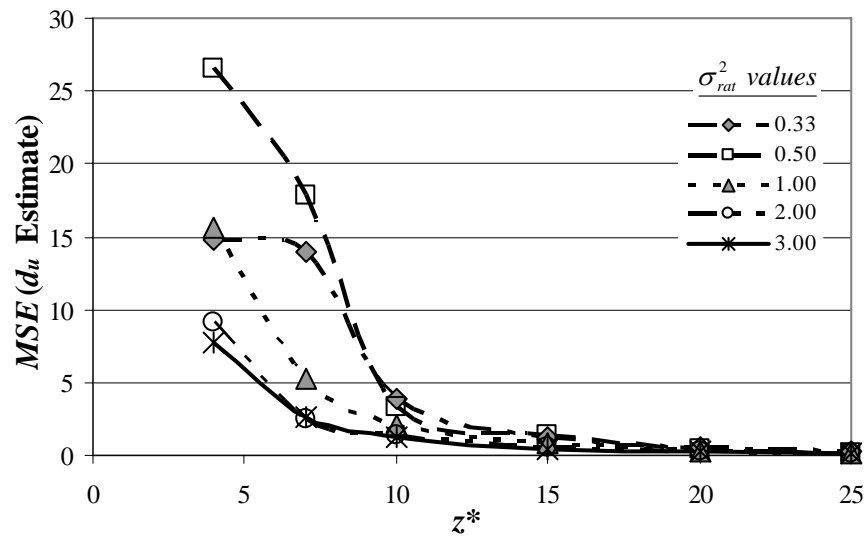
Not graphically illustrated in this dissertation was the impact of varying n_T on the both treatment effect representation *bias* and *MSE* measures. Consistent with the $\hat{\phi}$ results from Figure 40 as well as the hypotheses conclusions, the results were invariant to changing n_T when z^* was fixed.

Section 2.4.5 explored the relationships between z^*, d_u, and d^*, culminating with the formulaic expression shown in Equation (96). That expression, however, was predicated on conditions of variance equality in the two populations, where the comprehensive study evaluated ratios other than 1.0. Using a similar development process from that section, allowing for unequal sample size and variance ratios, the relationship between z^* and d^* is represented as

$$z^* = d^* \sqrt{\frac{(n_C + \phi n_T)(1-\phi)n_T}{n_C + n_T}} \left(\sqrt{\frac{(n_C + \phi n_T)(\sigma^2_C/\sigma^2_T) + (1-\phi)n_T}{(n_C + \phi n_T) + (\sigma^2_C/\sigma^2_T)(1-\phi)n_T}} \right) \quad (107)$$

Using the effective sample size relationships defined in Equations (33) and (34), this can be rewritten into a more concise form

$$z* = d^* \sqrt{\frac{n_C^* \, n_T^*}{N}} \left(\sqrt{\frac{n_C^* \, \sigma_{rat}^2 + n_T^*}{n_C^* + n_T^* \sigma_{rat}^2}} \right) \qquad (108)$$

where the term inside the parentheses equals 1.0 with population variance equality. This

relationship is necessary in support of the power analysis framework presented in the

next chapter.

3.3 Individual Level

      After analyzing hypotheses conclusions from the min *AIC* selection and the

accuracy and consistency of the chosen model's parameter estimates, attention now shifts

to the individual respondent in the treatment sample. With evidence supporting treatment

nonresponse, the population mixing proportion estimate can be translated to an

individual's likelihood of being a treatment nonresponder using Bayes' theorem,

Equation (79). Because these calculations depend on model selection as well as

parameter estimates, they are subject to the same concerns noted in previous sections:

varied composition of selected models, interpretations of *bias* and *MSE*, and bimodality.

      The individual classification error, being an averaged measure, has interpretation

challenges in bimodal conditions, where this measure is heavily dependent on the mixing

proportion estimate. Taking into account these concerns, Figure 46 illustrated findings

similar to other performance measures; variance ratios exceeding 1.0 comparatively had

the best results and each σ_{rat}^2 condition became increasingly error free as z^* increased.

Figure 46.
Graph: Average Individual Classification Error at $\phi = 0.10$, $n_T = 200$,
over z^* by σ^2_{rat}
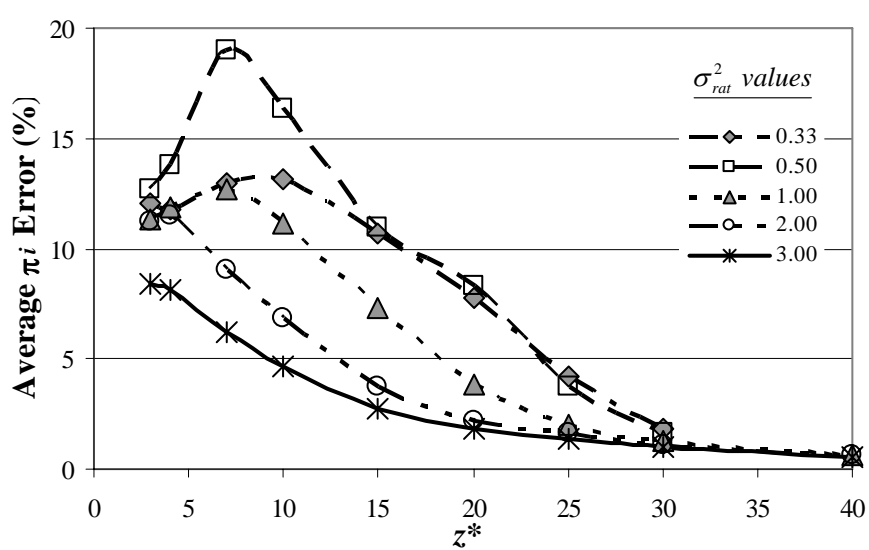
An alternative to individual classification error is a dichotomous classification of
an individual's group membership based on their posterior probability. These
classifications are aggregated to determine an overall correct classification rate for the
entire sample, shown in Figure 47, with fixed ϕ and n_T conditions different from the
previous figure.

Figure 47.
Graph: Correct Classification % at $\phi = 0.35$, $n_T = 100$ over z^* by σ^2_{rat}

104
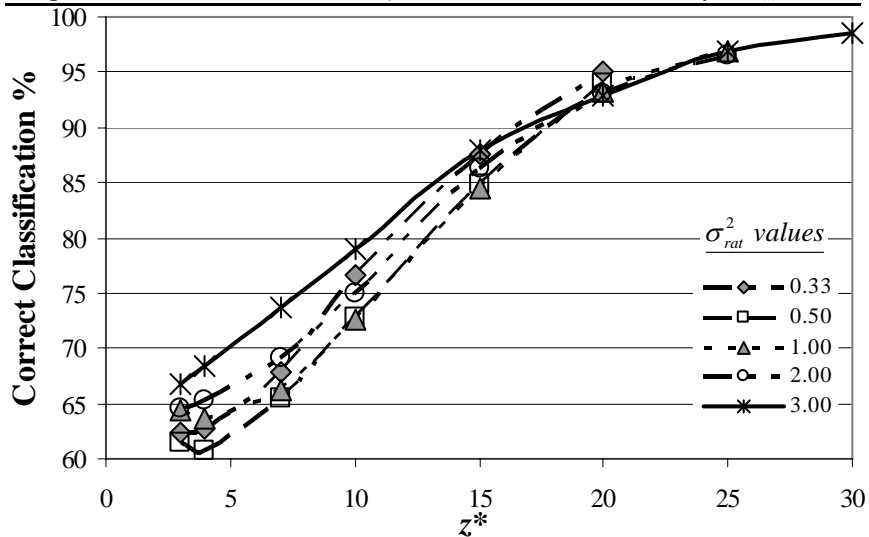
Each σ^2_{rat} condition approaches 100% correct classification as z^* increases, where perhaps resulting from the method of classification, this particular measure appears relatively robust to σ^2_{rat} changes. Performance of this same measure was also examined under varied n_T conditions fixing z^* for a single σ^2_{rat} value, shown in Figure 48.

Figure 48.
Graph: Correct Classification % at $\phi = 0.05$, $\sigma^2_{rat} = 0.50$ over z^* by n_T
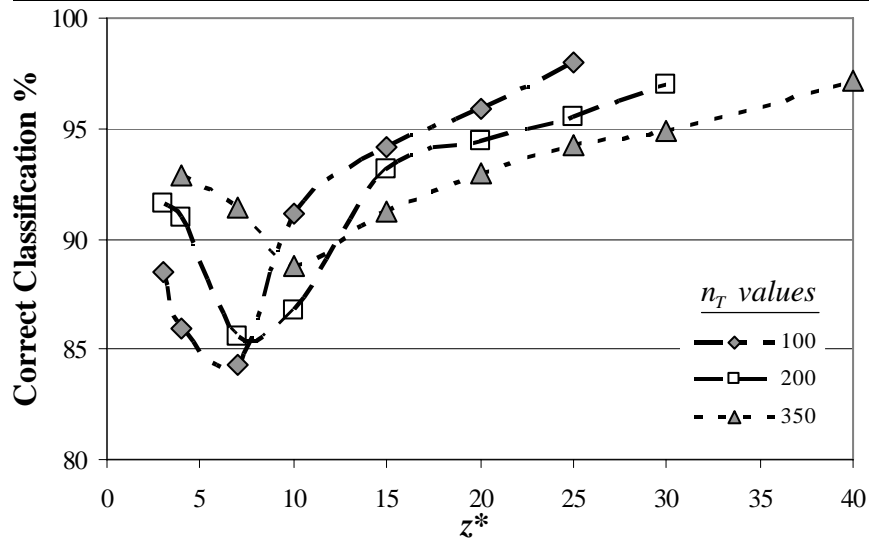
Unlike the population level measures and hypotheses conclusions, n_T demonstrates an effect when z^* was fixed. The effects, however, similar to those in correct model selection, are mixed where larger n_T values had improved performance at lower z^* values but comparatively worsened as z^* increased.

3.4 Results Summary

This section provides a synopsis and synthesis of all the findings from the pilot and comprehensive studies. First, the z^* composite parameter is central in relating to all of the performance measures supported by the following comments:

105

a. It demonstrates asymptotic properties for the correct treatment effect and mixture hypotheses conclusions, reaching 100% selection rates as $z*$ increases.

b. The findings in (a) are even more noteworthy as those asymptotic properties similarly hold in assessing the accuracy and consistency of parameter estimates of interest, as the *bias* and *MSE* of $\hat{d}_u$, $\hat{d}^*$, and $\hat{\phi}$ each move to 0 as $z*$ increases.

c. Exceptions to the asymptotic properties for correct model selection and the variance hypothesis condition occurred when the σ^2_{rat} was 1.0. In such conditions, the results asymptote around 83%, though still increase as z^* increases. All other population σ^2_{rat} conditions asymptote to a 100% selection rate.

d. The path towards convergence for each of the performance measures as a function of z^* differs conditioned on this ratio, where σ^2_{rat} values exceeding 1.0 consistently provided the best results.

e. Recapturing the individual class membership, higher $z*$ values resulted in increasingly error free classification, whether using an average individual error or overall sample correct classification measure. Unlike the other performance measures, the overall correct classification results seemed more robust to changes in the population variance ratio.

f. $Z*$ is moderately robust to departures in the effective sample size ratio from 1.0. This was graphically illustrated in Figure 19 as well as Equation (93), which in turn allowed equal sample sizes to be evaluated when transitioning

from the pilot study to the comprehensive study. The impact of this result is even though a researcher should increase the treatment group sample size if treatment nonresponse is suspected, the deviation from z^*_{max} is relatively small with limited effect on resulting hypothesis conclusions.

g. Perhaps the most significant finding for the key performance measures which are the three hypothesis conclusions and the *bias* and *MSE* of the reported parameter estimates $\hat{\phi}$, $\hat{d}_u$ and $\hat{d}^*$, these results were invariant to n_T when z^* was fixed.

h. For other performance measures such as correct model selection and overall correct classification, n_T affected results even when z^* was fixed. Findings were similar for each measure and varied as a function of z^* where increased n_T values comparatively improved results for smaller z^* values but comparatively worsened as z^* increased.

The population mixing proportion impacted hypotheses conclusions with fixed z^* values. Higher ϕ values generally resulted in the higher correct mixture hypotheses conclusions across all σ^2_{rat} values. In the highest ϕ values evaluated, correct mixture hypothesis conclusion rates decreased slightly, attributed to the large deviation in the effective sample size ratio from 1.0. The mixing proportion impact, however, was the opposite with respect to the variance equality hypothesis. Larger ϕ values decreased the correct variance hypothesis conclusion rate.

All of these results including the tables provided in Appendices 2-5 are predicated on the use of the *AIC* as the information criterion. While other information criteria could be the basis for model selection, based on the results the *AIC* is strongly recommended.

Aside from the reasons provided in the preceding chapters, the use of any sample size based parsimony penalty nullifies the valuable invariance properties of n_T with z^* in hypotheses conclusion results. Selection of another criterion is based on some assumption that its use will provide improved results. With particular focus on the mixture hypothesis, "improved" is a reduction in the false classification rates in Appendix 3 and increased selection rates in Appendix 4. Using total sample size in construction of the parsimony penalty, while false classifications decrease, the correct identification will worsen to some unknown extent. Additionally the mixed results in correct model selection and overall individual correct classification rates with n_T as a function of z^* are further exacerbated with a parsimony penalty that increases with N.

Chapter 4: Discussion

Against some 13 performance measures ranging from correct hypotheses conclusions to the quality of parameter estimation to recapturing group membership, this min *AIC* strategy is established as a viable technique for obtaining empirical evidence of treatment nonresponse in two sample designs. There are a number of other areas which require comment, however, in order to effectively transition this technique from a simulation environment to applied research. Further, advocating only this series of models is terribly shortsighted where the series of models in this study serve as a foundational block upon which to build and obtain empirical evidence for more complex research questions. Accordingly, this chapter was not titled "Discussion of Results", but more simply "Discussion" in a number of directions for this methodology to take root in the diverse world of applied research.

4.1 Transition to Applied Research

Transition to applied research greatly reduces the computational burden involved in this process with analysis of a single dataset. *Bias*, *MSE*, and correct selection rates in the experimental study are replaced with a reliance on the model selection process to provide correct hypothesis conclusions with accurate population estimates.

4.1.1 Defense of Parametric Distribution Specifications

Selection of appropriate parametric distributions is foundational from model development to model selection. Unlike empirical research which operates under controlled conditions, the onus is on the researcher for the selection and defense of

parametric distribution specifications for the control and treatment populations. This is a more stringent requirement than statistical tests of population means presented in section 1.1, yet is more flexible than fixed effects ANOVA, for example, which requires normality and variance equality. Assisting in this regard for the control population, a researcher can view a histogram and summary statistics from the control sample. Researchers may feel restricted focusing exclusively on only common parametric distributions, but there is a vast selection of available choices (Leemis, 2008).

Identification of a parametric distribution for the treatment population is more difficult. Selection of a distribution characterizing the treatment population is recommended, but is not required, to be the same distribution as the control population. Evaluation of summary statistics or viewing a histogram from a treatment sample, however, is not informative and potentially misleading, where the presence of some unknown proportion of nonrespondents affect the results. To illustrate the difficulty in assessing the underlying parametric distribution from the treatment sample, the following mixture probability density functions (PDF) shown in Figure 49 are from a normally distributed treatment population whose mean and variance values varied in each plot. Each plot had a 0.20 proportion of treatment nonresponse from a normally distributed control population. Overlaid on each plot in dashed lines is a normal PDF to illustrating skewness, kurtotic, or even bimodal properties.

Figure 49.
Graph: Various Probability Density Function Mixtures from Normal Distributions
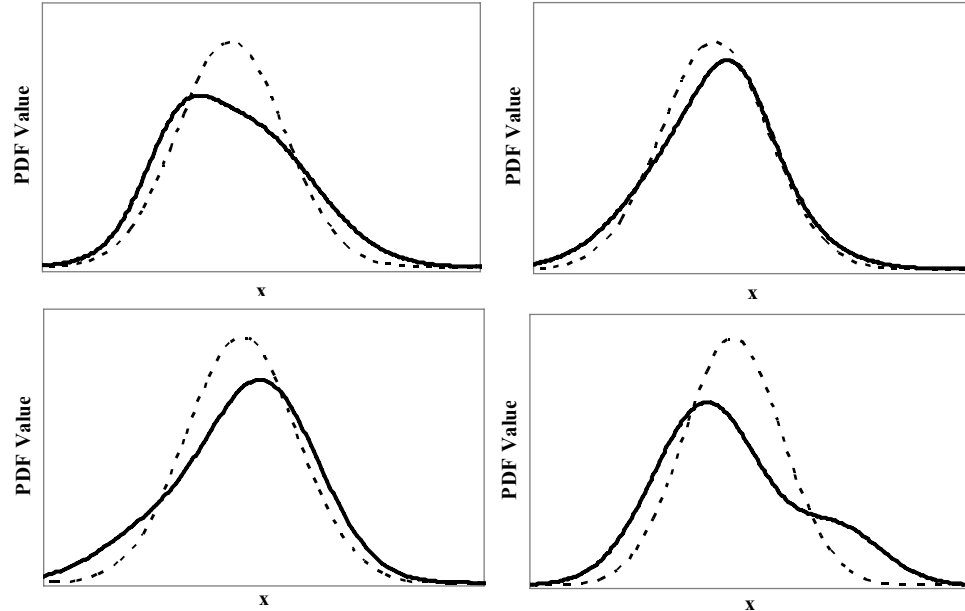
Until now, the information based approach has not required any probabilistic

inferences and Type I error control upon which to base hypothesis conclusions. A min

AIC strategy selects the most reasonable representation of the data comparatively from

among a set of models. Despite having the best comparative fit, this fails to serve as an

absolute fit measure, available in other modeling procedures such as R^2 in regression and

the *SRMR* in structural equation modeling. Because this technique has no measure of

absolute fit for reference and selection of appropriate parametric distributions underpins

the entire process, non-parametric distributional tests should be conducted as a degree of

support for these choices, where the Kolomogorov-Smirnov (K-S) test (D'Agostino,

1986) is recommended.

Prior to presentation of the K-S test, it should be noted that statisticians have

concerns with this test, or more generally any distributional tests. Their concerns are

two-fold involving the critical test statistic determined as a function of *N*. First, in case of

small sample sizes, such tests are not powerful enough to reject an incorrect parametric

111

distribution. With very large sample sizes, the test statistic becomes so small that any trivial distributional deviation results in a rejection of the null hypothesis. While these are valid concerns, like the *bias* and *MSE* measures presented last chapter, there does not seem to be another alternative available. The bottom line is some type of justification is required for these parametric distribution choices. Whether the presentation of questionable evidence in the form of a K-S test or operating under the unsupported assumption those are the correct population distributions is better remains a fair and open debate.

The K-S test compares the empirical distribution function (EDF) from a sample of data against a fully specified hypothesized distribution's cumulative density function. Preference for this test in lieu of other nonparametric tests is calculation of its test statistic and critical values do not depend on the hypothesized distribution. Further, the K-S test is an exact test, and shown to be more powerful than goodness of fit tests, which are highly dependent on sample size and the number of bins. A disadvantage to this particular test, however, is the critical values become less reliable when distributional parameters are estimated from the data. For that reason, for certain common distributions, the Anderson-Darling or Shapiro-Wilk (for normality) tests have been shown to be more powerful (Sheskin, 2007). These tests, however, do not have corresponding critical test statistic values to support evaluation of a mixture distribution CDF, the desired result of the min AIC selection process represented in the treatment sample. Adjustments to the K-S test, such a two-stage K-S variant introduced by Khamis (2000) have been shown to increase the power of the test, demonstrated for common distributions such as the normal and exponential, but do not support mixture CDFs.

Because there are two data samples, two K-S tests are necessary, where the first test on the control group sample can occur prior to any model selection process. Observing the sample histogram for the control group, a researcher can select a parametric distribution representative of the control population or perform a data transformation in order to obtain to a more well suited structure. Any data transformations, however, must be consistently applied to both data samples in order for treatment nonrespondents in the treatment group sample to retain the same distributional properties. To construct the EDF, reorder the control group sample, $\vec{x}_C$, from smallest to largest, where the EDF is represented as

$$\text{EDF}_C \mid x_{Ci} = \frac{i}{n_C} \tag{109}$$

for each value in the data sample where i corresponds to the index number in the reordered data vector. The hypothesis being evaluated using the EDF is

$Ho : F(x_C) = F^*(x_C)$ - the data follows the specified distribution
$Ha : F(x_C) \neq F^*(x_C)$ - the data does not follow the specified distribution

Distributional parameter estimates are determined by using the summary statistics shown previously in Equations (36) and (37) for a normal distribution specification. Of interest is the maximum distance between the EDF step function and hypothesized distributional CDF expressed in the following formula

$$D_C = \max\left[F^*(\hat{\boldsymbol{\theta}} \mid x_{Ci}) - \frac{i-1}{n_C}, \frac{i}{n_C} - F^*(\hat{\boldsymbol{\theta}} \mid x_{Ci}) \right] \tag{110}$$

A researcher would reject the null hypothesis when

$$D_C > D_{\text{Crit}} = \frac{1.22}{\sqrt{n_C}} \tag{111}$$

113

which is the approximate critical value of the two-sided hypothesis test at $\alpha = 0.10$ (Conover, 1999). Use of this data sample omits possible treatment nonrespondents population parameters estimates may differ slightly from those resulting from a min *AIC* selection. Nonetheless, there is sufficient information to achieve the goal of this test; some evidence supporting the parametric distribution selection for the control population.

Given the difficulty in identifying a treatment population distribution from Figure 49, using the same parametric distribution successfully defended in the initial K-S test, construction of the EDF for the treatment group sample follows in the manner where

$$\text{EDF}_{\text{T}} \mid x_{\text{T}i} = \frac{i}{n_{\text{T}}} \tag{112}$$

The distribution to be evaluated under this hypothesis test, $F^*(x_{\text{T}})$, is returned with parameter estimates as part of the min *AIC* model selection of the form

$$F^*(x_{\text{T}}) = \hat{\phi} F_{\text{C}}(\hat{\boldsymbol{\theta}}) + (1 - \hat{\phi}) F_{\text{T}}(\hat{\boldsymbol{\theta}}) \tag{113}$$

If both distributions are normally distributed, this coincides with the set of models presented in Figure 3 and results presented in Chapter 3. Following the min *AIC* model selection, therefore, the hypothesis

$Ho : F(x_T) = F^*(x_T)$ - the data follows the min *AIC* specified distribution
$Ha : F(x_T) \neq F^*(x_T)$ - the data does not follow the min *AIC* specified distribution

is evaluated comparing the following test statistic

$$D_{\text{T}} = \max\left[F^*(\hat{\boldsymbol{\theta}} \mid x_{\text{T}i}) - \frac{i-1}{n_{\text{T}}}, \frac{i}{n_{\text{T}}} - F^*(\hat{\boldsymbol{\theta}} \mid x_{\text{T}i}) \right] \tag{114}$$

against the critical value shown in equation (109). A *failure to reject* conclusion provides a degree of evidence affirming use of the same parametric distribution for each

114

population. Conversely, a rejection of the null hypothesis indicates the same parametric distribution for each population is not empirically supported, where a different parametric distribution for the treatment population can be explored. Specification of another population distribution will, however, change the number and construction of models as subsequently presented in section 4.2.1.

A broader type of comparative fit assessment might also be used as a degree of support for the parametric distributions choices. A researcher may evaluate a larger set of competing models positing various distributions, still using a min *AIC* selection as the mechanism for model selection.

4.1.2. Software Transition

Transitioning the software built to conduct this experimental study to support applied research also requires modifications. The addition of the K-S tests must be included within the software. The biggest addition is a broader selection of parametric distributions, consistent for both populations and subsequently expanded to allow different parametric distribution combinations. At times, transformations of data approximating normality are not tenable or not preferred to retain the original data scale and interpretation. The number and construct of models evaluated under this min *AIC* methodology are determined from the parametric distribution specifications, where results from normally distributed populations presented in Chapter 3 do not generalize to other parametric distributions.

Regardless of the distribution specifications, the use of an optimization algorithm is still required, where default tolerance and maximum number of search iterations can be

increased from the experimental study. Creation of reasonable starting values can follow

a similar process utilizing sample summary statistics. Time savings garnered by

analyzing a single set of data allow a more rigorous method ensuring a global minimum

of the -2ln(L) function to be implemented. Due to the large number of trials and

experimental conditions, the current optimization routine for selected models is

conducted up to 15 times, stopping at any point if the respective model criteria are

satisfied. As an alternative, the optimization routine would run a fixed number of

iterations, 25 for example, retaining the likelihood function value and corresponding

parameter estimates. From among the 25 likelihood function values, select the smallest

confirming the model criteria are satisfied. This provides, via a comparative assessment

of optimization routine results on the same data differing only in starting values, greater

confidence of a global solution.

Finally, the results of this normal distributions study should be incorporated into

the software, assisting the researcher from a power analysis perspective and in sample

size planning for their particular study.

4.1.3 A Power Analysis Framework

In an applied setting, it seems unreasonable not to entertain empirical evidence as

part of the model selection process. In fact, researchers consistently use empirical results

of a wholly theoretically based model, whether a mixture model specification or more

classic IID representation. Because this methodology advocates an empirically based

model selection, a degree of power should be afforded to a researcher. There is no

comparative analog with an *a priori* model selection or technique which results in

116

inferential statements of model parameters. Providing the researcher a power analysis

framework in model selection reflects a combination of both frequentist and Bayesian-

like approaches. A frequentist approach makes no conditions upon parameter estimates

in terms of being influenced with *a priori* knowledge. However, a predetermined model

selection applies *a priori* knowledge, making it Bayesian-like in nature. In many ways,

this is worse where a purely Bayesian approach allows empirical evidence to influence

results; *a priori* model selection does not.

This does not suggest that *a priori* knowledge and theory should not be part of

model selection. On the contrary, it is critically important, but must be bolstered by

empirical evidence. The fusion of a researcher's knowledge and empirical evidence

from this study can be performed both pre-model and post-model selection for a series of

models evaluating treatment nonresponse. Power analysis does not eliminate the need for

the K-S tests or other defense of the parametric distributional choices.

Prior to evaluating any models, a researcher positing normal distributions for the

populations can get an estimate of the likelihood of obtaining empirical evidence

supporting the existence of treatment non-response. From the experimental study results

presented in Appendix 4, a researcher can obtain

$$\hat{\Pr}(\text{Correct}_{\text{Mix}} \mid \hat{z}^*, \hat{\phi}, n_{\text{T}}, \hat{\sigma}_{\text{rat}}^2) \qquad (115)$$

based on their belief of population conditions and available sample sizes. Likelihood

estimates can be interpolated, linearly or by some other method, for population conditions

that do not directly coincide with table values (e.g., a ϕ estimate of 0.25 or a z^* value of

12). As noted in Chapter 3, the lowest likelihoods of obtaining a correct mixture

hypothesis conclusion occurred at σ^2_{rat} of 1.00 or 0.50, which could be used as a worst case approximation.

EXAMPLE: A researcher conducted a two group study consisting each of 150 respondents. He/she posits both populations subscribe to normal distributions that result in the following conditions

$$\left\{ \hat{z}* = 15, \hat{\phi} = 0.20, \hat{\sigma}^2_{rat} = 1.00 \right\}$$

and is interested in the likelihood of obtaining the correct mixture hypothesis conclusion. First, a researcher can compute the expected biases for d_u and Cohen's d by using summary statistics in lieu of utilizing a maximum likelihood framework with Equations (67) and (68). Using the tables in Appendix 4,

$$\hat{\Pr}(\text{Correct}_{\text{Mix}} \mid \hat{z}* = 15, \hat{\phi} = 0.20, n_{\text{T}} = 100, \hat{\sigma}^2_{rat} = 1.00) = .832$$

and

$$\hat{\Pr}(\text{Correct}_{\text{Mix}} \mid \hat{z}* = 15, \hat{\phi} = 0.20, n_{\text{T}} = 200, \hat{\sigma}^2_{rat} = 1.00) = .716$$

where an interpolation provides a likelihood estimate of 0.767 from this min *AIC* strategy.

A researcher might also utilize the experimental study results in sample size planning prior to conducting a particular study to ensure a minimally acceptable likelihood for a correct mixture hypothesis conclusion using the Appendix values if normal distribution specifications are made for each population. Because sample sizes are a component in z^* calculation, this is a more complex and iterative process. The iterative process begins with estimates of the population conditions and a reasonable $\hat{z}*$ value where the following relationship can be used

$$\hat{n}_{\mathrm{T}} \approx \frac{(\hat{z}*)^2 \left(\hat{\sigma}_{\mathrm{C}}^2 (1 - \hat{\phi}) + \hat{\sigma}_{\mathrm{T}}^2 (a + \hat{\phi}) \right)}{\left(\hat{d}_u \right)^2 (a + \hat{\phi})(1 - \hat{\phi})} \tag{116}$$

where a is the sample size ratio defined in Equation (58). With this estimate, a researcher

now has all the elements for Equation (115) to refer to Appendix 4.

EXAMPLE: A researcher is planning a study and wants to determine the

minimum sample sizes required for a 90% likelihood of empirical evidence supporting

treatment nonresponse from the following conditions

$$\left\{ \hat{\mu}_{\mathrm{C}} = 5, \, \hat{\sigma}_{\mathrm{C}}^2 = 20, \, \hat{\mu}_{\mathrm{C}} = 10, \, \hat{\sigma}_{\mathrm{T}}^2 = 10, \, \hat{\phi} = 0.15, \, a = 1.00 \right\}$$

To begin, the researcher selects an arbitrary z^* value of 15, and using Equation (116)

results in a n_{T} requirement of

$$\hat{n}_{\mathrm{T}} \approx \frac{(15)^2 \left(20 * (1 - 0.15) + 10 * (1 + 0.15) \right)}{\left(|5 - 10| \right)^2 (1 + 0.15)(1 - 0.15)} \approx 262.4 = 263$$

With the treatment sample size calculated, the researcher can interpolate a value shown in

the table below

Table 13.
Corresponding Extracts from Appendix 4 for Mixture Hypothesis Evidence
for Sample Size Planning Example

ϕ	n_t value		
	200		350
0.10	89.0	$(n_T = 263)$	87.4
	$- \left(\hat{\phi} = 0.15 \right) \rightarrow$		
0.20	96.2		96.4

resulting in an estimated likelihood of 0.922, slightly above the desired power. As a

result, the researcher can decrease the sample size, conscious of the fact this will also

lower z^* value. Lowering the sample size to 200, using equation (63) reduces $z*$ from 15

to 13.09. Interpolation from Appendix 4 leads to a new estimated likelihood 0.83, an

overcorrection in the sample size reduction. After a second iteration, 240 respondents per

sample, given the desired sample size ratio, are necessary to achieve the desired power

obtaining a correct mixture hypothesis conclusion.

A researcher can also evaluate their *a priori* knowledge conditioned upon the

results of the model selection. Assuming a mixture model was chosen, such a result is

only possible when

<div align="center">
Mixture Model Chosen | Treatment Group had nonresponse
or
Mixture Model Chosen | Treatment Group had zero nonresponse
</div>

Using the information regarding false classification tables in Appendix 3 with Equation

(90), this probability can be estimated. For a correct mixture hypothesis conclusion when

treatment nonresponse is present, the estimated probability becomes

$$\frac{\hat{\Pr}(\text{Correct}_{\text{Mix}} \mid \hat{z}^*_{\phi=\hat{\phi}}{}^*, \hat{\phi}, n_{\text{T}}, \hat{\sigma}^2_{\text{rat}})}{\hat{\Pr}(\text{Correct}_{\text{Mix}} \mid \hat{z}^*_{\phi=\hat{\phi}}{}^*, \hat{\phi}, n_{\text{T}}, \hat{\sigma}^2_{\text{rat}}) + \beta^*_{Mix} \mid \hat{z}^*_{\phi=0}{}^*, \hat{\phi} = 0, n_{\text{T}}, \hat{\sigma}^2_{\text{rat}}} \tag{117}$$

EXAMPLE. A researcher conducts a two group study consisting each of 200

respondents, positing normal distributions for both populations with the following

estimates

$$\left\{ \hat{z}^* = 20, \ \hat{\phi} = 0.10, \ \hat{\sigma}^2_{\text{rat}} = 1.00 \right\}$$

Using Appendix 4, the likelihood of obtaining empirical evidence supporting treatment

nonresponse is 0.844. If indeed there was zero treatment nonresponse, the z* value

would be slightly increased due to the denominator changes in Equation (63), where

$\hat{z}^* \mid \phi = 0.0$ equals 20.1. The likelihood of a false mixture selection under these conditions

can be found in Appendix 3 as 0.052. If the model selection process results in the

<div align="center">120</div>

mixture hypothesis conclusion supporting treatment nonresponse, then using Equation

(117), the *a priori* information results in

$$\frac{.844}{.844 + .052} = 0.942$$

which is a significant improvement in the likelihood from Appendix 4.

Conversely, the complement to this model selection result can also be developed:

the case when a no mixture model was incorrectly selected. This likelihood is estimated

$$\frac{1 - \hat{\Pr}(\text{Correct}_{\text{Mix}} \mid \hat{z}^*_{\phi=\hat{\phi}}*, \hat{\phi}, n_{\text{T}}, \hat{\sigma}^2_{\text{rat}})}{\left(1 - \hat{\Pr}(\text{Correct}_{\text{Mix}} \mid \hat{z}^*_{\phi=\hat{\phi}}*, \hat{\phi}, n_{\text{T}}, \hat{\sigma}^2_{\text{rat}})\right) + (1 - \beta^*_{\text{Mix}} \mid \hat{z}^*_{\phi=0}*, \hat{\phi} = 0, n_{\text{T}}, \hat{\sigma}^2_{\text{rat}})} \qquad (118)$$

EXAMPLE. Using the same experimental conditions from the previous example,

the min *AIC* model selection process selects a model that provides a $\hat{\phi}$ of 0. Using

Equation (118), the likelihood this was an incorrect hypothesis conclusion is

$$\frac{(1 - .844)}{(1 - .844) + (1 - .052)} = 0.141$$

There are simpler and more general planning tools available to a researcher

unsure of population distributional parameters in order to determine z^* required for power

analysis. The basis for these tools is to maximize z^* as function of a, the sample size

ratio, σ^2_{rat}, and ϕ. Further, Equation (64) illustrates z^*'s suitability for any parametric

distribution specification. Maximizing z^* capitalizes on its asymptotic properties with

regard to correct hypothesis conclusions, among other performance measures,

demonstrated in the experimental study of Normal distributions. These properties remain

under consideration of different parametric distributions as shown in the next section.

Now, with beliefs regarding σ^2_{rat} and ϕ, a researcher can determine the optimal sample

size ratio, $a_{\max}$, to achieve $z^*_{\max}$ using

$$a_{\max} = \sqrt{\sigma_{\text{rat}}^2}\,(1-\phi)-\phi \qquad (119)$$

Valid sample size ratios, however, must be such that $a_{\max} > 0$. Notice this formula is independent of N and any distributional parameter estimates. This formula does not determine the value of $z^*{}_{\max}$, but the position in 3-dimensional space, $\{\sigma_{\text{rat}}^2,\ \phi,\ a_{\max}\}$, where this value exists.

In many research situations, however, including the experimental study, $a_{\max}$ may not be utilized where instead equal sample sizes, $a = 1$, is a common choice. In such cases, the z^* value is some percentage below the $z^*{}_{\max}$ value. This percentage or deviation from $z^*{}_{\max}$ is determined

$$\frac{z^*{}_{a=1}}{z^*{}_{a_{\max}}} = \sqrt{\frac{(1+\sqrt{\sigma_{\text{rat}}^2})^2\,(1-\phi^2)}{2(\sqrt{\sigma_{\text{rat}}^2}-\phi\sqrt{\sigma_{\text{rat}}^2}+1+\phi)}} \qquad (120)$$

is a generalization of Equation (93) and likewise is independent of N and distribution parameters. A graphical representation of this deviation, presented in Figure 50, is a 3-D extension of Figure 19 additionally varying σ_{rat}^2. The larger deviations for σ_{rat}^2 conditions less than 1 under equal sample sizes might explain, in part, performance differences in variance ratios equidistant from 1 noted in Chapter 3. Researchers using a sample size between 1 and $a_{\max}$ can use this figure to estimate the deviation from $z^*{}_{\max}$ as a result of the sample size ratio selection.

Figure 50.

Graph: Deviation from z^*_{max} when using $a = 1$ by σ^2_{rat} and ϕ
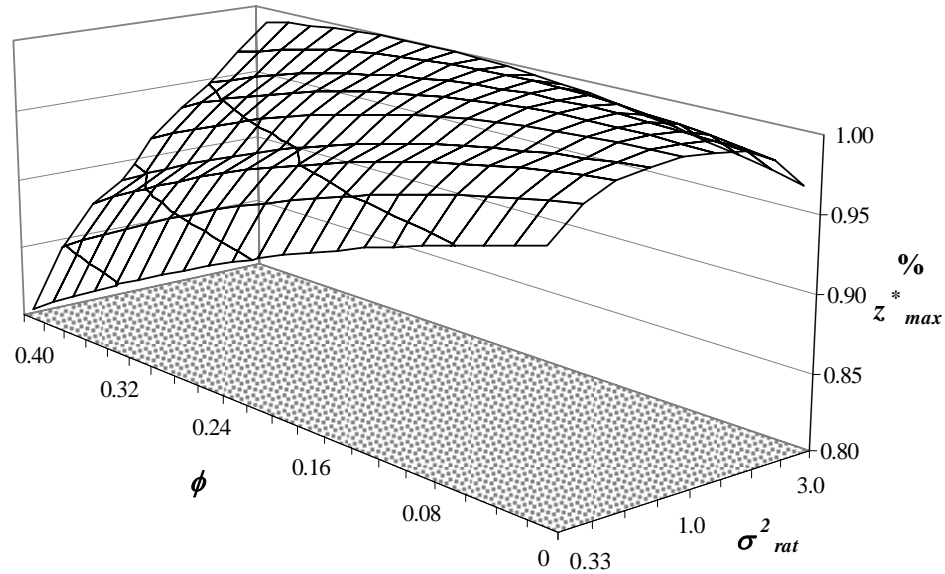
4.2 Other Parametric Distributions

A flexibility of this methodology is its ability to specify almost any parametric distribution. Both Titterington (1985) and Grun (2002) indicate mixtures of these types are mathematically tractable for most distributions, yet the performance of a min *AIC* strategy under a two sample design has not been evaluated. The set of research questions, presented in Section 1.5 have an overarching level: distributional. To illustrate the suitability of other parametric distributions and differences across performance measures due to distributional specifications, populations from a Poisson distribution will be considered. *Z**, the mixing proportion, and treatment sample sizes are extensible to other parametric distributions and are retained as control parameters enabling comparisons with the normal population distributions.

4.2.1 Poisson Distribution

Unlike the normal distribution, the Poisson distribution is a single parameter distribution, characterized by λ requiring discrete data. The probability mass function (PMF) for this distribution is

$$P(X = x) = pois(x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{121}$$

whose mean and variance are the same

$$E(\mathbf{X}) = V(\mathbf{X}) = \lambda \tag{122}$$

Unlike the normal distribution example used to this point, the researcher now assumes that

$$\mathbf{X}_C = x_{C1}, x_{C2}, \ldots x_{Cn_C} \rightarrow f_C(\lambda) \quad \text{for } i = 1, 2, \ldots, n_C \tag{123}$$

represents the control group and treatment group is characterized by either

$$\mathbf{X}_T = x_{T1}, x_{T2}, \ldots x_{Tn_T} \rightarrow f_T(\lambda)$$

$$or \tag{124}$$

$$\mathbf{X}_T = x_{T1}, x_{T2}, \ldots x_{Tn_T} \rightarrow \phi f_C(\lambda) + (1 - \phi) f_T(\lambda) \quad \text{for } j = 1, 2, \ldots n_T$$

for models posited without and with treatment nonresponse. Development of these models under a maximum likelihood framework proceeds in the same fashion as section 1.4 using Equation (29). Because this is a single parameter distribution, however, the series of comparative models are fewer, illustrated in Figure 51.

Figure 51.
Figure 51.
Model Representations with Mixing Proportion with Poisson Distributions

MIXTURE MODEL
$0 < \phi < 1$

TRADITIONAL MODELS
$\phi = 0$

Model #1
$\lambda_C \neq \lambda_T$
2 Parameters

Model #2
$\lambda_C = \lambda_T$
1 Parameter

Model #3
$\lambda_C \neq \lambda_T$
3 Parameters
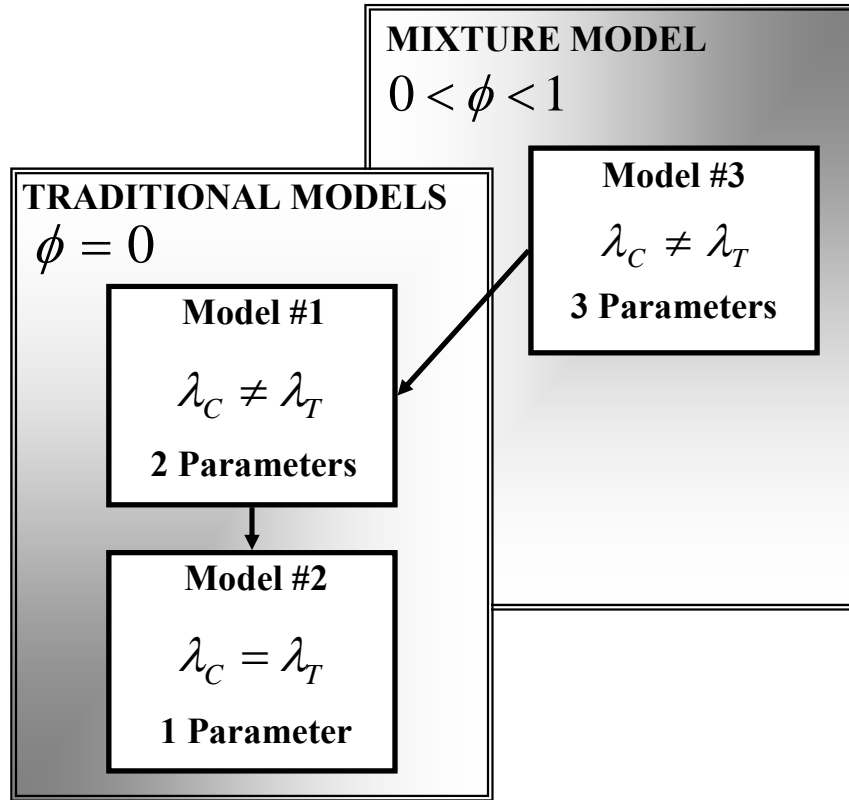
Referencing Figure 3, normal distribution Models #2, #3, and #6 cannot be constructed as

mean and variance constraints across populations must be consistent. There is not a

separate variance hypothesis. Because of Poisson distribution specifications, selection of

a model with a treatment effect also results in unequal variances, where a selection of no

treatment effect indicates variance equality. Model selection corresponding to a correct

mixture hypothesis and treatment effect conclusion is shown below.

Table 14.
Correct Hypothesis Result by Model Selection for Poisson Distributions

Population Conditions*	Correct Hypothesis Result		
ϕ	Correct Model	for ϕ	for $E(\mathbf{X})$
$\phi = 0$	Model #1	Model #1, 2	Model #1, 3
$0 < \phi < 1$	Model #3	Model #3	Model #1, 3

* Note: All population conditions operated with a population treatment effect.

Without available software to support an evaluation, code was developed in Gauss using the same validation, technical decisions and specifications outlined in section 2.1. Only the mixture model representation required an optimization algorithm where the others utilize mathematically proven results. For Poisson Model #1, the minimum value of the -2ln(L) function occurs at

$$\hat{\lambda}_{\mathrm{C}} = \bar{x}_{\mathrm{C}} \qquad\qquad \hat{\lambda}_{\mathrm{T}} = \bar{x}_{\mathrm{T}} \qquad\qquad (125)$$

Poisson Model #2 requires a single parameter estimate whose minimum -2ln(L) value occurs at

$$\hat{\lambda} = \hat{\lambda}_{\mathrm{C}} = \hat{\lambda}_{\mathrm{T}} = \bar{y} \qquad\qquad (126)$$

where y is the concatenated samples. Poisson model #3 has 3 parameters where the optimization algorithm has the following constraint

$$0 < \hat{\phi} < 1 \qquad\qquad (127)$$

where

$$-2\ln(L), \left\{ \hat{\lambda}_{\mathrm{C}}, \hat{\lambda}_{\mathrm{T}}, \hat{\phi} \right\} \qquad\qquad (128)$$

are returned. Starting value generation, number of optimization attempts, and model fit requirements established by its hierarchical relationship were similarly applied.

4.2.1.1 Experimental Conditions

Experimental conditions evaluated include both an increase and decrease of λ between the control and treatment populations. To facilitate comparison to the normal population distributions results, the λ values were set at 1.00 and 2.00, which is a variance ratio of 0.5 and 2.0 depending on which is assigned as the control population. A corresponding graphical depiction of the population PMFs is provided in figure 52.

Figure 52.
Graph: Poisson PMFs for Experimental Conditions where $\sigma^2_{rat} = 2.0$ and 0.50
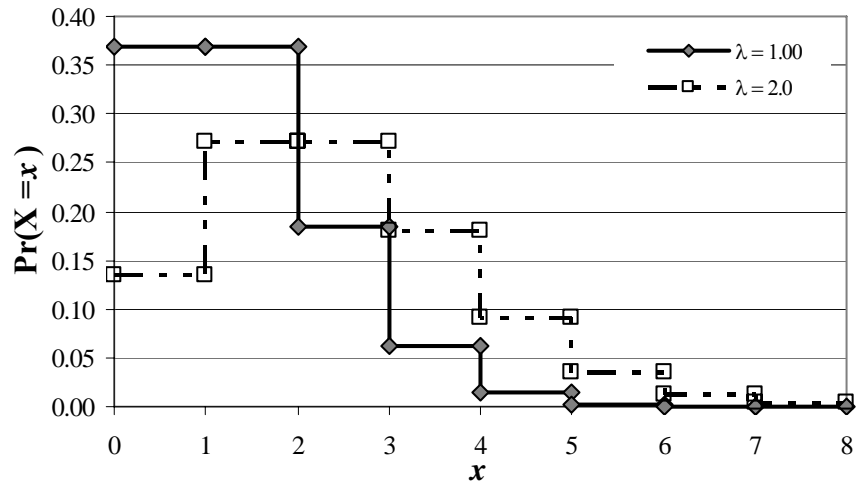
Because the Poisson is a single parameter distribution, the standard deviation is

subsequently fixed by the λ term, so z^* can not be systematically varied as done in the

previous study. The ϕ and n_T values, however, are varied in the same levels as the

normal distribution variance ratio equivalents. Tables 15 and 16 provide the

experimental conditions in addition to the associated z^* and d^* values.

Table 15.
Empirical Conditions: Poisson Distribution $\lambda_C = 1.00$, $\lambda_T = 2.00$, $n_C = n_T$

ϕ	0.00	0.05	0.10	0.20	0.35	0.50
n_C	z^*	z^*	z^*	z^*	z^*	z^*
100	5.77	5.72	5.65	5.48	5.12	4.63
200	8.16	8.09	7.99	7.75	7.24	6.55
350	10.80	10.70	10.57	10.25	9.57	8.66
d^*	0.82	0.82	0.83	0.85	0.87	0.89

Table 16.
Empirical Conditions: Poisson Distribution $\lambda_C = 2.00$, $\lambda_T = 1.00$, $n_C = n_T$

ϕ	0.00	0.05	0.10	0.20	0.35	0.50
n_C	z^*	z^*	z^*	z^*	z^*	z^*
100	5.77	5.81	5.84	5.86	5.75	5.48
200	8.16	8.22	8.26	8.28	8.14	7.75
350	10.80	10.88	10.93	10.95	10.76	10.25
d^*	0.82	0.81	0.80	0.79	0.77	0.76

Another series of experimental conditions compared normal distribution results of

variance ratios of 0.33 and 3.00, where the λ values selected were 2.25 and 0.75.

Population PMFs are provided in Figure 53 with the experimental conditions provided in

Tables 17 and 18.

Figure 53.

Graph: Poisson PMFs for Experimental Conditions where $\sigma^2_{rat} = 3.00$ and 0.33
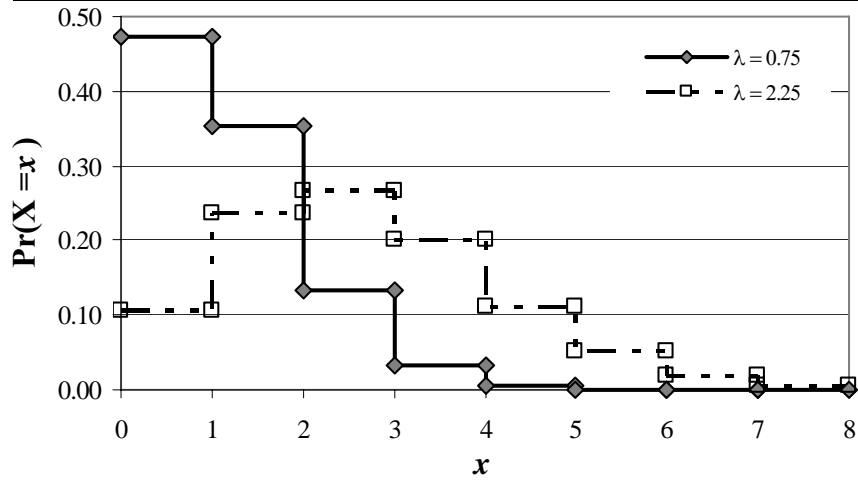
Table 17.

Empirical Conditions: Poisson Distribution $\lambda_C = 0.75$, $\lambda_T = 2.25$, $n_C = n_T$

ϕ	0.00	0.10	0.20	0.35
n_C	z^*	z^*	z^*	z^*
100	8.66	8.41	8.09	7.48
200	12.25	11.89	11.44	10.58
d^*	1.22	1.26	1.29	1.35

Table 18.

Empirical Conditions: Poisson Distribution $\lambda_C = 2.25$, $\lambda_T = 0.75$, $n_C = n_T$

ϕ	0.00	0.10	0.20	0.35
n_C	z^*	z^*	z^*	z^*
100	8.66	8.84	8.94	8.93
200	12.25	12.50	12.65	12.63
d^*	1.22	1.20	1.17	1.13

<u>4.2.1.2 Results</u>

Results from the Poisson distribution study are not as detailed as the Normal distributions study where the primary purpose is to illustrate differences resulting from parametric distribution specification.   Unlike the Normal distribution study, the optimization algorithm did not converge for every trial. Despite increasing the converging attempts to 20 per trial, convergence rates ranged between 70-100%. Convergence rates improved as difference between the population λs increased and with increased ϕ values. This result dispels the complement to the false belief from the Normal distribution study that a failed convergence implies model misspecification. No analysis was conducted pinpointing the predominant cause of failed convergence, whether the *SQP* algorithm, errors resulting from matrix inversion, not satisfying model hierarchical relationships, or some other factor.   Completion of the study allowed a distributional comparison under similar z^*, ϕ, n_T, and σ^2_rat experimental conditions. A qualitative summary of selected performance measures are provided in Table 19 along with more particular assessments for two subgroups created with σ^2_rat values less than and greater than 1. At the model level, no observable distributional differences were noted for the treatment effect hypothesis as selection rates were 100% in nearly all the conditions examined. For correct model selection, and more generally the preponderance of performance measures, the Poisson distribution specification outperformed the Normal distribution whose λ_rat and σ^2_rat values were below 1.0. Selection of the correct model yielded mixed results illustrated in the comparison of the distributional specifications, shown in Figure 54.

Table 19.
Distributional Performance Measure Comparison with the Same ϕ, n_T, z^*
and σ^2_{rat} Conditions

Performance Measure	Poisson Comparison to Normal		
	$\lambda_{rat} = \sigma^2_{rat} < 1$	$\lambda_{rat} = \sigma^2_{rat} > 1$	Overall
Model Level*			
Correct Model Selection	Better	Worse	Mixed
Correct ϕ Hypothesis Result	Mixed	Worse	Worse
Population Level			
d^*_{bias}	Better	Equal	Better
$MSE(\hat{d}^*)$	Better	Equal	Better
$\hat{\phi}_{bias}$	Worse	Worse	Worse
$MSE(\hat{\phi})$	Better	Worse	Mixed
Individual Level			
$\bar{\pi}_{error}$	Better	Worse	Mixed
$\bar{\pi}_{\%Class}$	Better	Worse	Mixed

Figure 54.
Graphs: Correct Model Comparisons {$\sigma^2_{rat} = 0.50$, $n_T = 100$} and {$\sigma^2_{rat} = 3.00$, $n_T = 200$}
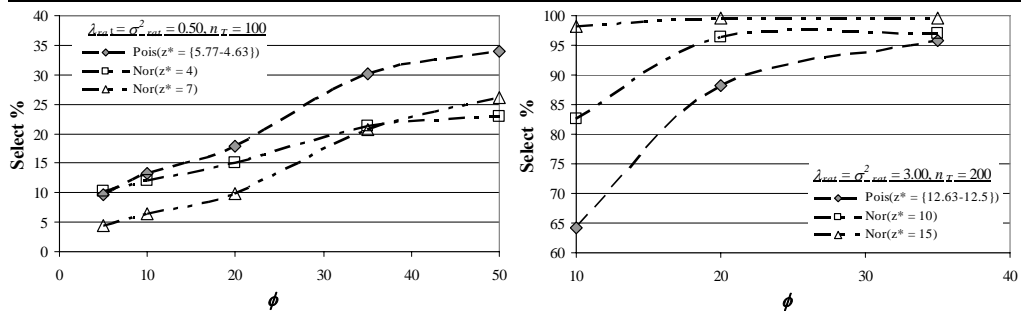
Examining hypothesis conclusions, the Normal distribution more frequently

arrived at the correct mixture hypothesis conclusion, shown in Figure 55.   Possible

explanations for this finding include the flexibility of the Normal distribution, the

coarseness of data required in Poisson models, or the change in the distributional shape

commensurate with a λ change.

Figure 55.

Graphs: Mix Hypothesis Comparisons $\{\sigma^2_{rat} = 0.33, \ n_T = 100\}$ and $\{\sigma^2_{rat} = 2.00, n_T = 350\}$
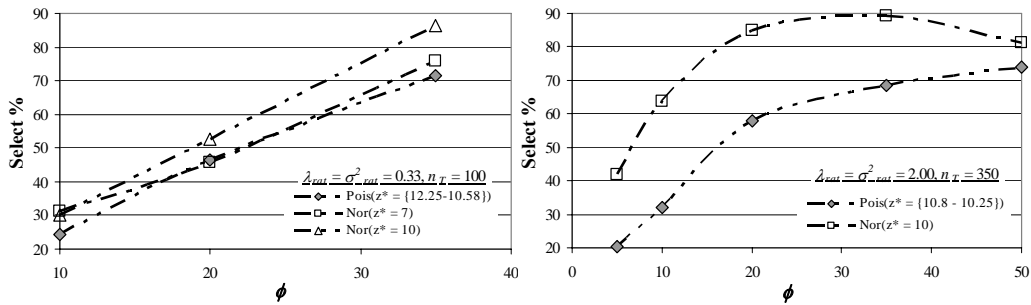
For population level measures, the ability to recapture the unstandardized

treatment effect, d_u, was not done as these values were different between studies. The

Poisson distribution comparatively performed better in recapturing the population

standardized treatment effect in terms of a lower *bias* and *MSE*. The result could also be

attributed to the Normal distribution's flexibility, in this case serving as a detriment.

With a larger of number of parameters simultaneously estimated to calculate a composite

measure, $\hat{d}^*$, a larger *bias* and larger *MSE* should be an expected result, shown in Figures

56 and 57.

Figure 56.

Graphs: Bias *d** Estimate Comparisons $\{\sigma^2_{rat} = 0.50, \ n_T = 200\}$ and $\{\sigma^2_{rat} = 2.00, \ n_T = 100\}$
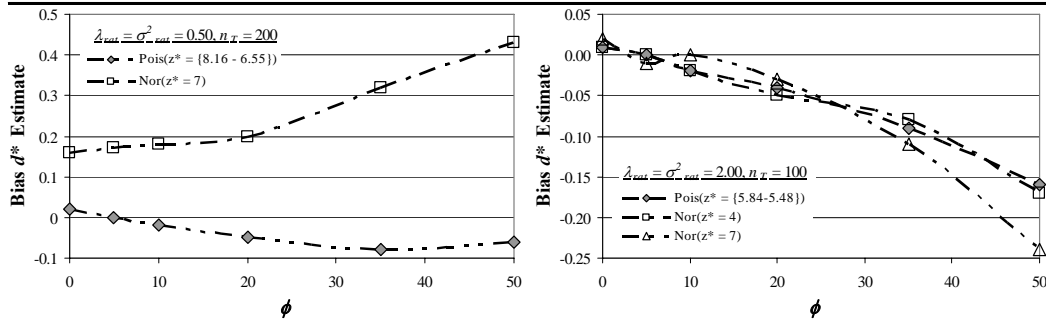
Figure 57.

Graphs: *MSE d** Estimate Comparisons { $\sigma^2_{rat} = 0.50$, $n_T = 100$ } and { $\sigma^2_{rat} = 2.00$, $n_T = 200$ }
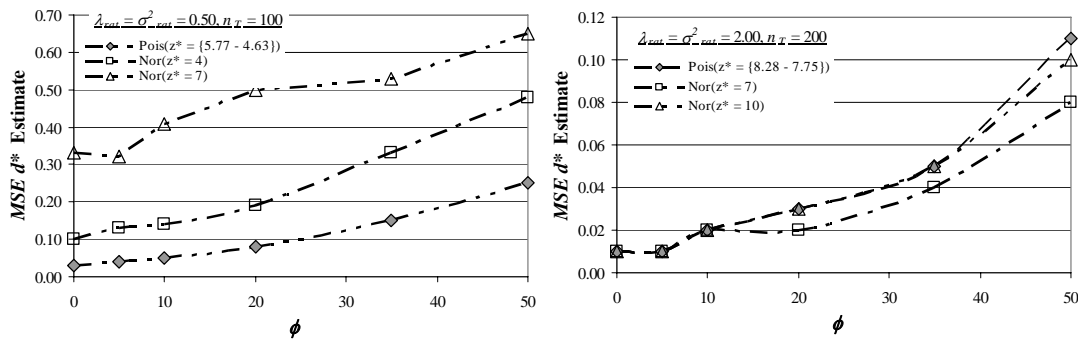
The population mixing proportion estimate, $\hat{\phi}$, unlike $\hat{d}^*$, is not a composite measure. Perhaps due to this condition, the Normal distribution provided less biased results than the Poisson distribution as shown in Figure 58. Comparison of $\hat{\phi}$ *MSE* values returned mixed results due largely to the poor performance of Normal distribution at a variance ratio of 0.50.

Figure 58.

Graphs: Bias ϕ Estimate Comparisons { $\sigma^2_{rat} = 0.33$, $n_T = 200$ } and { $\sigma^2_{rat} = 3.00$, $n_T = 100$ }
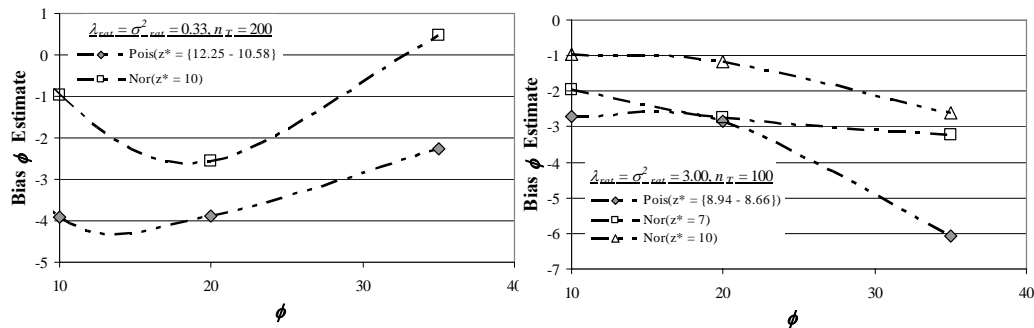
At the individual level, comparative evaluation of both performance results were mixed, where the percentage of correct classification is presented in Figure 59.

Figure 59.

Graphs: Individual Classification Comparisons $\{\sigma^2_{rat} = 0.33, n_T = 200\}$ and $\{\sigma^2_{rat} = 2.00, n_T = 200\}$
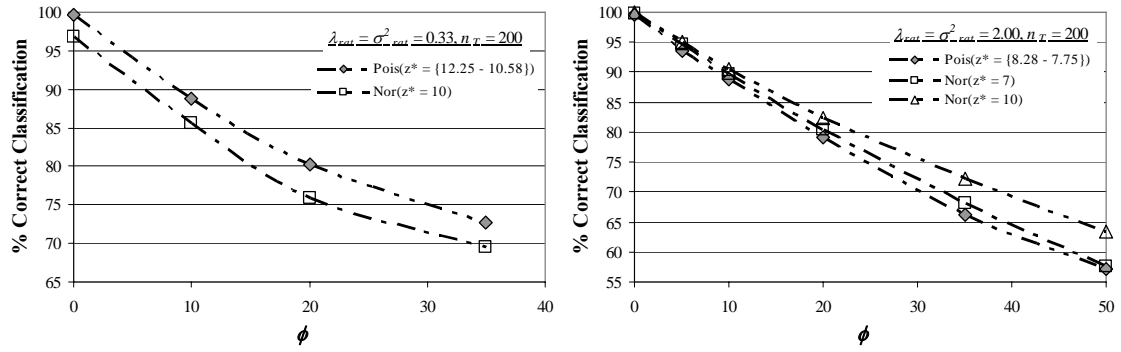
4.2.2 Near Equivalent Probability Representations:

The previous section indicated Poisson distribution specifications are

comparatively more difficult to obtain the correct mixture hypothesis conclusion.

Despite using the same controlled parameters, $\{z^*, \phi, n_T, \text{ and } \sigma^2_{rat}\}$, differences could be

attributed to the distributional shape, where the Poisson data representations explored

were highly skewed and kurtotic.   Another explanation could be the coarseness of

measurement, where the Poisson distributions used discrete data. Making the conditions

similar across both distributional specifications, for larger values of λ, the Normal

distribution becomes an excellent approximation of data from a Poisson population

(Devore, 2000) where

$$Pois(\lambda \mid x) \approx Nor(\mu, \sigma \mid x) \tag{129}$$

with the following distributional parameter relationships

$$\lambda = \mu = \sigma^2 \tag{130}$$

Using these relationships in order to evaluate the min *AIC* strategy with data equally well

characterized by either distribution, larger λ values of 10 and 15 were selected, with the

distribution specific PMFs or PDFs shown in Figure 60.

133

Figure 60.
Graph: Poisson PMF / Normal PDF Comparison $\{\lambda_C = 10, \lambda_T = 15\}$
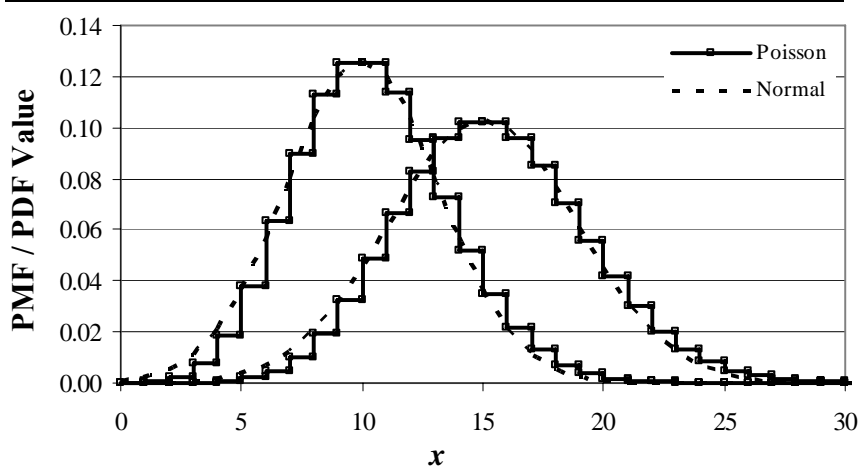
From these population conditions, 500 trials were conducted with the Normal distribution

specification models presented in section 1.4 and the Poisson distribution specification

models introduced in the previous section. Both an increase and decreasing in the

treatment effect were evaluated using equal sample sizes where $n_T = 200$ and a ϕ of 0.20.

For an increase in the population treatment effect, $\{\lambda_C = 10, \lambda_T = 15\}$, the z^* value was

13.6 with a corresponding variance ratio of 0.67. A histogram comparing the model

selection results for each distributional specification is shown in Figure 61 below.

Figure 61.
Graph: Model Selection Percentages $\{n_T = 200, \phi = 0.10, \lambda_C = 10, \lambda_T = 15\}$
for Poisson and Normal Distribution Equivalents
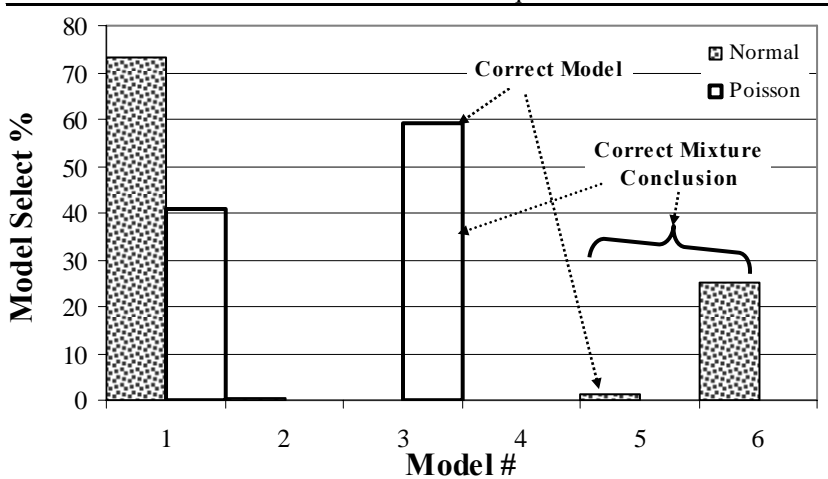
Comparison of the correct model selection is startling, where even the mixture hypothesis

conclusions are significantly improved when specifying Poisson distributions, differing

from the results observed in the last section. Poisson is a more parsimonious distribution

where the normal distribution requires estimation of twice as many parameters to assess

the same hypotheses. This is, in essence, an application of Occam's razor in regards to

parametric distribution selection. A full comparison of results, now including a σ^2_{rat} of

1.5 is provided in Table 20.

Table 20.
Performance Results Near Equivalent Probability Representations
(Normal and Poisson Distributions)

Performance Measure	$\{\lambda_C = 10, \lambda_T = 15\}$		$\{\lambda_C = 15, \lambda_T = 10\}$	
	Normal	Poisson	Normal	Poisson
Model Level*				
Correct Model Selection	1.2%	59.8%	58.4%	73.4%
Correct ϕ Hypothesis Result	26.0%	59.8%	61.2%	73.4%
Correct σ^2 Hypothesis Result	74.4%	n/a	76.4%	n/a
Population Level				
$\hat{\phi}_{bias}$	-3.39	-1.89	-2.54	-.88
$MSE(\hat{\phi})$	147.3	59.2	54.2	45.9
d^*_{bias}	-.03	-.02	-.04	-.01
$MSE(\hat{d}^*)$	.09	.02	.03	.02
$\hat{d}_{u\,bias}$	-.12	-.05	-.13	-.03
$MSE(\hat{d}_u)$	.54	.27	.21	.20
Individual Level				
$\bar{\pi}_{error}$	11.38%	6.58%	6.20%	5.51%
$\bar{\pi}_{\%\,Class}$	87.90%	89.94%	90.75%	90.82%

* Note: Each distribution specification resulted in 100% correct hypothesis conclusions regarding
population means.

For every performance measure, Poisson distribution specifications outperformed an

equally well fitting Normal distribution specification.

4.3 Methodological Extensions

Using the framework developed in Chapter 1 bolstered by the experimental study

results in Chapter 3, a number of more complex research questions can be accommodated within this framework, where introductory development of a few examples is provided within this section. Those examples include multiple treatments, multiple responses within a treatment, *a priori* consideration of covariates, and most promisingly multivariate responses within a latent framework. By no means is this an exhaustive list of all possible extensions. Normal distributions for the control and treatment populations are again used as examples in this section

Inherently supporting more complex research questions are more complex models, which in a comparative model framework create a combinatoric expansion of less parsimonious models to evaluate. In these cases, evaluation of *all* possible models is impractical, where the researcher must defend their choice of a reasonable and sufficient subset of models. One reduction option for normal distribution specifications is to no longer evaluate the hypothesis for variance equality, making each population variance parameter freely estimated. As a result of this decision, the set of models presented in Figure 3 is reduced to

Figure 62.
Normal Distributions Model Reduction Option
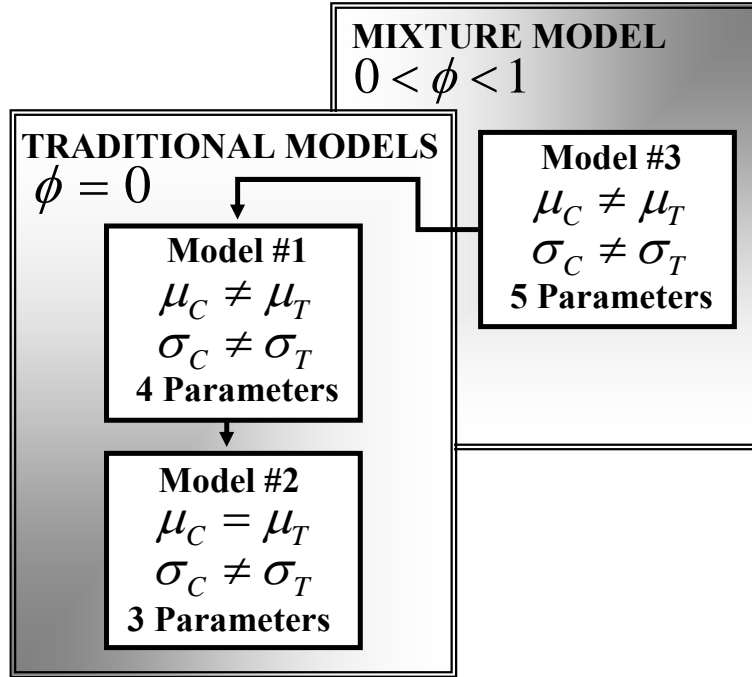
where Equation (29) remains the basis for these constructed models. This set of models

still simultaneously evaluates mixture and treatment effect hypotheses. Most

importantly, the results and formulaic use of Appendices 2-4 can still be used for this

model set, though Appendix 5 is no longer valid. Unchanged is the requirement of

researcher input for the population variance ratio to extract the appropriate likelihood

value. Presentation of the univariate extensions use this reduced set of models thereby

omitting evaluation of variance equality.

4.3.1 Multiple Treatments

The first research extension evaluates two different treatments, treatment *A* and

B, with the same control group, where two comparative frameworks can be utilized. The

simpler framework is separate analyses between the control group and each treatment due

to independence in the administration of the treatments, shown in Figure 63.

Figure 63.
Model Composition Assessing Multiple Treatments (Option #1)
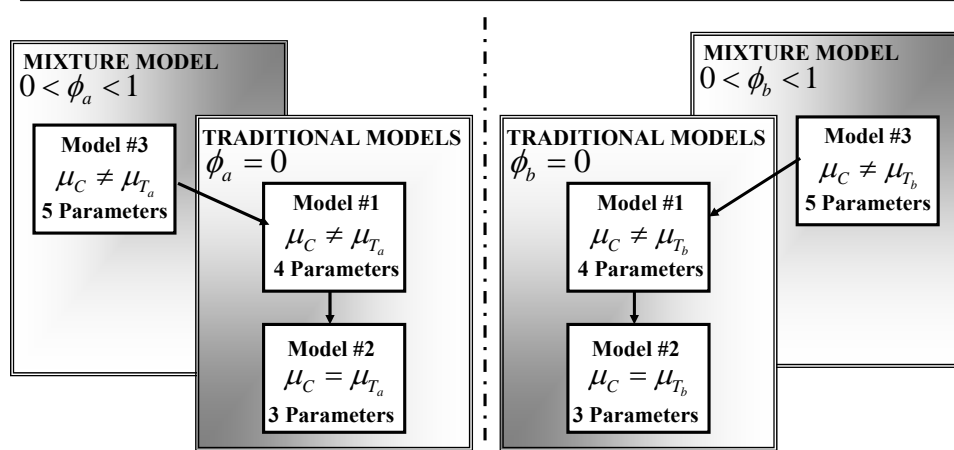
Beyond mixture hypothesis conclusions, a secondary interest is differences in

the two treatments, whether in degree of treatment nonresponse or differences in

population means. Upon model selections, population estimates can be reported for each

treatment. $\hat{\phi}_a$, $\hat{\phi}_b$, or $\left(\hat{\phi}_a - \hat{\phi}_b \right)$ do not subscribe to a normal distribution, where reporting

of confidence intervals or inferential statement for these estimates is not readily available.

Equation (64) can be used to generate a p-value with regard to the comparison of

population mean estimates.

Resulting from independent analysis for each treatment, interpretation and

reporting of control population parameters becomes challenging. Under any number of

model selection scenarios, estimates for the control population parameters will be

different with separate analysis. While the differences may be quite small, one option is

to report an average from both analyses. A more rigorous option and second framework

is to consider both treatments simultaneously with the common control group through a

single model selection process. Where the previous option totaled six models and two

min *AIC* selections, this option is a single min *AIC* selection of 13 models. Composition

of these models is presented in Figure 63, where because of its complexity, the display of

hierarchical relationships and number of parameters estimated for each model has been

omitted.

Figure 64.
Model Composition Assessing Multiple Treatments (Option #2)
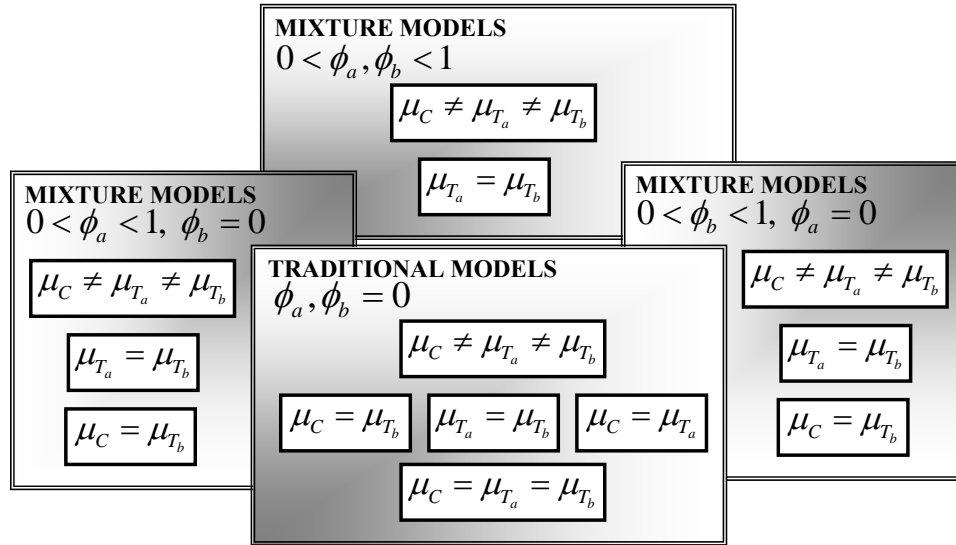
Not only does a single model selection process eliminate conflicting control population

estimates, assessing differences in treatment means follows as part of model selection. If

a researcher was also interested in variance equality assessments, the addition of this

hypothesis would require an evaluation of 35 different model specifications, a near 3 fold

increase. Empirical evidence assessing the homogeneity of $\hat{\phi}_a$ and $\hat{\phi}_b$ can also be

evaluated in a comparative model process with new models added to figure 63, but is not

presented. K-S tests should be conducted for the control sample initially and post model

selection for each separate treatment sample as defense for the parametric distribution

choices.

4.3.2 Multiple Responses to a Single Treatment

There is an obvious difference in interpretation between nonresponse to a treatment and multiple responses to a treatment, where distinction between these outcomes is not afforded in single sample experimental designs. With the presence of a control group, empirical evidence for and delineation of these outcome becomes tenable, represented in a series of five models presented in figure 64 which considers two (multiple) different responses.

Figure 65.
Model Composition Assessing Multiple Responses to a Single Treatment
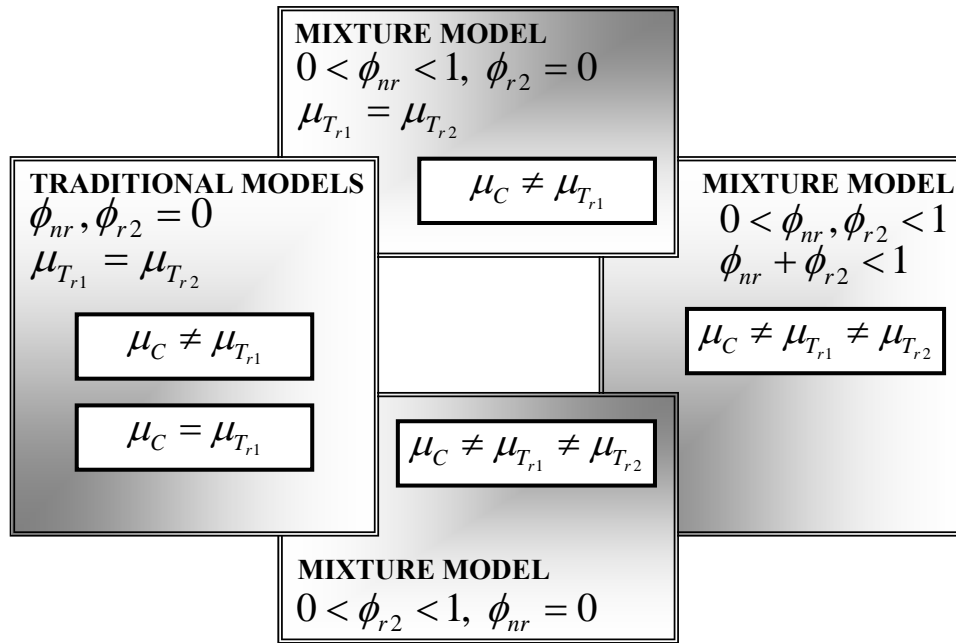
Three of the models are the reduced set of normal distributional models previously presented where the constraints

$$\{\mu_{T_{r1}} = \mu_{T_{r2}}, \phi_{r2} = 0\} \tag{131}$$

indicate the treatment response is represented by a single distributional structure.

The remaining two models have been added as layers to support evaluating a treatment response as a mixture of two distributional responses. These model

specifications posit two different treatment responses, where unlike Equations (2) or (26) without and with treatment response, the treatment sample is represented as the following distributional form

$$\mathbf{X}_T = x_{T1}, x_{T2},...x_{Tn_T} \sim \phi_{nr} f_C(\mathbf{\theta}) + \phi_{r2} f_{T2}(\mathbf{\theta}) + (1 - \phi_{nr} - \phi_{r2}) f_{T1}(\mathbf{\theta})$$

$$\text{for } j = 1,2,...n_T$$

(132)

A normal distribution is specified for the 2nd treatment response.   The model at the bottom of Figure 64 with ϕ_{nr} constrained to 0 indicates a treatment sample without treatment nonresponse characterized by a mixture of two different responses while the far right model of Figure 64 has both treatment nonresponse and two distinct responses.

While the treatment effect hypothesis remains generally unchanged, the hypothesis regarding mixtures and a homogeneous *(ID)* population now creates a number of alternatives where

$$Ho_1 : \phi_{nr}, \phi_{r2} = 0 \qquad \text{- Homogeneous } \textit{(ID)} \text{ population}$$
$$Ha_{1a} : \phi_{nr} > 0, \phi_{r2} = 0 \text{ - Treatment Nonresponse Exists, Single Treatment}$$
$$\text{Response}$$
$$Ha_{1b} : \phi_{nr} = 0, \phi_{r2} > 0 \text{ - Zero Treatment Nonresponse, Multiple (2) Treatment}$$
$$\text{Responses Exist}$$
$$Ha_{1c} : \phi_{nr}, \phi_{r2} > 0 \qquad \text{- Treatment Nonresponse and Multiple (2) Treatment}$$
$$\text{Responses Exist}$$

Adding the variance equality hypothesis increases the set of models to 13. K-S tests should be conducted in the same manner where this test can accommodate the most complex representation where the sample is a mixture of three population distributions.

4.3.3 Mixture Extension Considering Covariate *A Priori*

Often, supplemental information for respondents is available affording greater depth in analysis. Should a mixture model be advocated from this process, post hoc analysis of posterior probabilities of group membership, responders and nonresponders, based on these covariates provide tremendous insight. In the comparative model strategies presented assessing treatment nonresponse and the subsequent research extensions, determination of a population mixing proportion was based only on an observed response.

Consideration of covariates in an *a priori* fashion, however, is a fundamentally different issue and quite challenging. Such a decision implies with the same treatment response, respondents with differing covariate information will have different likelihoods of being classified a nonrespondent. This changes the two sample construct maintained throughout to one focused on different subgroups of the treatment sample. Such a change enables each subpopulation to have different proportions of treatment nonresponse for the same treatment. Figure 66 provides an example where the interest in the treatment group is divided by gender and handedness, which also requires separation of the control sample to sub-populations by the same covariates.

Figure 66.
Mixture Extension Considering Covariates *A Priori*

CONSIDERATION OF A PRIORI COVARIATE INFORMATION
(Sub-populations)

(Example: Gender & Handedness)
1 Population becomes $K = 4$ Sub-populations

Sub-population #1: Male, Right-Handed	Sub-population #2: Male, Left-Handed	Sub-population #3: Female, Right-Handed	Sub-population #4: Female, Left-Handed
$f_{C_1} = ?$ $f_{T_1} = ?$ $\phi_1 = ?$	$f_{C_2} = ?$ $f_{T_2} = ?$ $\phi_2 = ?$	$f_{C_3} = ?$ $f_{T_3} = ?$ $\phi_3 = ?$	$f_{C_4} = ?$ $f_{T_4} = ?$ $\phi_4 = ?$
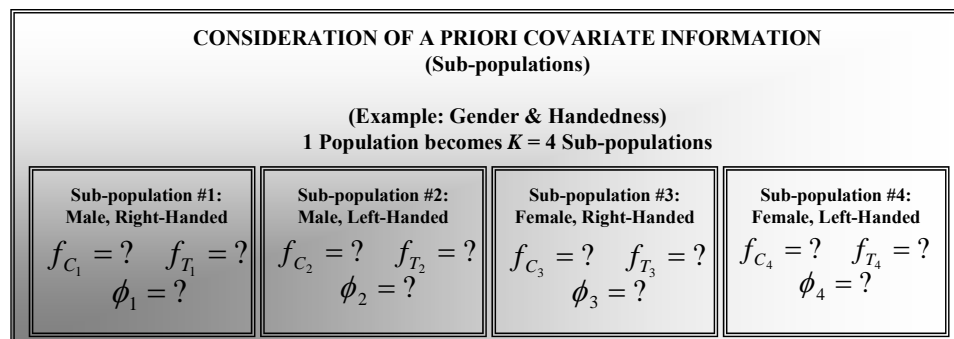
Unlike ANOVA, distributional normality is not required nor is equality of variance, but the researcher is now required to specify control and treatment parametric distributions for each subpopulation. Division of these samples makes the choice of parametric distributions more difficult and reduces the overall information *(power)* available to the min *AIC* selection, conducted *K* times for each *K* subpopulation shown in Figure 67.

Figure 67.
A Priori Analysis Process per Sub-population
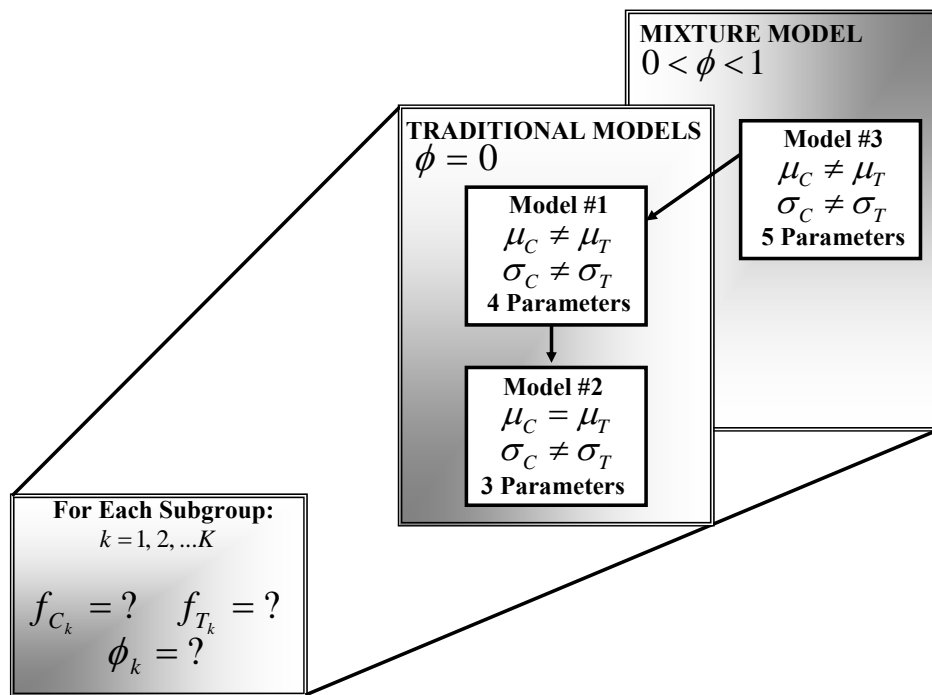
Separate K-S tests are required for each subpopulation. Comparison of subpopulation treatment means can be accomplished with traditional statistical tests with an appropriate multiple comparison correction. There is not any type of statistical test available to compare $\hat{\phi}$ across subpopulations though competing models could be assessed in evaluating homogeneity of $\hat{\phi}_k$ in selected pairs.

The inclusion of covariate information in analysis is recommended post model selection from the set of models represented in Figure 62 or the more extensive set in Figure 3 only if treatment nonresponse is supported. In such cases, no longer is dependent variable the treatment response; instead, it is replaced by either the posterior probability of being a treatment nonresponder using Equation (79) or the dichotomous treatment class assignment measure per respondent from Equation (82).

4.3.4 Multivariate Responses under a Latent Construct

The last two sample extension example occurs when a researcher has multiple measured outcomes. This section is not a complete development, only a sufficient introduction. Proposed is a latent variable or factor approach where the measured outcomes collectively fix a factor in location and scale (Hancock, 2004). This approach fundamentally differs from multivariate analysis of variance (MANOVA) which provides empirical evidence on composites of measured variables. Hancock (2003) and Cole (1993) provided greater clarity on the distinctions between MANOVA and latent variable methods. Currently available software, such as MIXFIT, offers a MANOVA solution evaluating mixtures within a single sample (McLachlan, 1998).

Experimental design with latent variables has a much shorter history, yet a number of software applications such as LISREL, EQS, and MPlus were developed in response to the increased popularity of these methods. The research interest in latent means differences changes the analysis from strictly covariance structures to augmented moment matrices or structured means modeling (SMM) (Sörbom, 1974), facilitating hypothesis evaluation directly at the construct level. With consideration of treatment

nonresponse at the factor level, the method of empirical evidence similarly transitions

from inferential statements on model parameter estimates to an omnibus, information

based selection process from competing models. Of the SMM modeling software

programs mentioned, only MPlus accommodates mixtures within a particular sample.

Particular to MPlus, the most recent 5.1 version allows constraints to be specified

across populations characterized by multivariate samples. There are, however, three

distinctions which prevent this program from supporting the proposed methodology:

a. While MPlus allows intercept and factor loadings constraints across groups

and specification of a number of latent classes in the treatment group, it does

not allow the properties of one of the latent classes to be fixed, crucial to

assessing treatment nonresponse. Shown subsequently in Figure 68, this

coincides with one of the latent classes from the treatment sample to be

$$\boldsymbol{\theta}_{F2,\,Class1} = \boldsymbol{\theta}_{Control} \sim \mathrm{Nor}(0, \sigma_C) \tag{133}$$

where the control group latent factor, assumed normally distributed, is

centered at 0.

b. Discussed in section 2.1.3, successful convergence takes advantage of the

hierarchical nesting in a series of models, requiring an improvement on a

$-2\ln(L)$ value. MPlus does not incorporate results from other model

representations in parameter estimation providing an extra degree of insurance

against local optimum solutions.

c. Most importantly, MPlus relies on robust statistics resulting from a Satorra-

Bentler correction, shown to provide more accurate inferential statements on

model parameters for data departures from multivariate normality (Curran,

West, & Finch, 1992; Satorra, 1994, 2003). These robust statistics do not provide inferences on the mixing proportion estimate, where instead a difference in latent means is used as a proxy in defense of its existence. Calculation of these robust statistics are a function of the data structure, the model implied augmented covariance structure, and the Jacobian matrix. This becomes confusing with a multiple model assessment where there would be different corrections per model specification altering the -2ln*(L)* values used in *AIC* construction. Multivariate normality, however, is exclusively a data condition which is unchanged for all these models.   A min *AIC* strategy comparing various model representations as an omnibus selection replaces utilization of robust statistics defending results intended for a single model specification. Robust statistics may be subsequently used assessing particular factor loadings and intercepts on the selected model.

As an example, consider an experiment involving a series of $r = 3$ measured indicators where

$$\mathbf{X}_C = \vec{\mathbf{x}}_{C1}, \vec{\mathbf{x}}_{C2}, ... \vec{\mathbf{x}}_{Cn_C} \sim \mathrm{Nor}(\vec{\mathbf{\mu}}_C, \mathbf{\Sigma}_C) \quad \text{for } i = 1,2....,n_C \tag{134}$$

represents a n_C by 3 matrix of responses for a control sample of size n_C characterized by a multivariate normal distribution. Independence is assumed between respondents where after accounting for the effects of the single latent variable, the responses are unrelated. For the treatment group, let

$$\mathbf{X}_T = \vec{\mathbf{x}}_{T1}, \vec{\mathbf{x}}_{T2}, ..., \vec{\mathbf{x}}_{Tn_T} \sim Nor(\vec{\mathbf{\mu}}_T, \mathbf{\Sigma}_T) \quad \text{for } j = 1,2,..,n_T$$

$$or \tag{135}$$

$$\sim \phi \, \mathrm{Nor}(\vec{\mathbf{\mu}}_C, \mathbf{\Sigma}_C) + (1 - \phi) \, \mathrm{Nor}(\vec{\mathbf{\mu}}_T, \mathbf{\Sigma}_T)$$

represent a n_T by 3 matrix of responses for a treatment sample of size n_T characterized by

a multivariate normal distribution. The same assumptions apply for the treatment sample,

where a representation of this latent structure is illustrated in Figure 68.

Figure 68.
Mixture Extension for Multivariate Responses Under a Single Latent Variable Construct
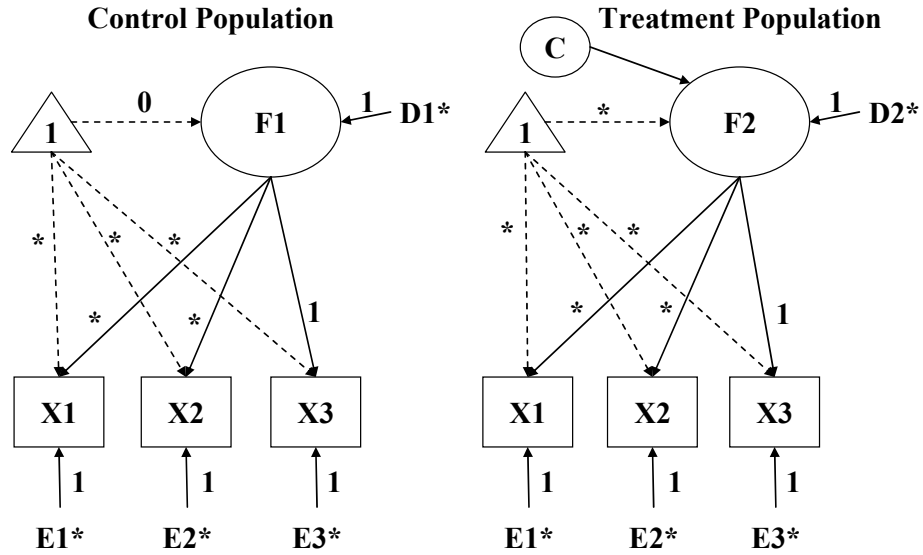
With traditional structured means models, where the number of latent classes, C, is one,

each latent variable is assumed to follow a normal distribution represented by its

particular sample

$$F1 \sim \text{Nor}(\mu_{F1}, \sigma_{F1})$$
$$F2 \sim \text{Nor}(\mu_{F2}, \sigma_{F2})$$

(136)

When evaluating model representations which posit treatment nonresponse, C becomes 2,

represented as

$$F1 \sim \text{Nor}(\mu_{F1}, \sigma_{F1})$$
$$F2_{\text{Class1}} \sim \text{Nor}(\mu_{F1}, \sigma_{F1})$$
$$F2_{\text{Class2}} \sim \text{Nor}(\mu_{F2}, \sigma_{F2})$$

(137)

The two structural components presented in Figure 68 are considered a single model for parameter estimation. Some comments relative to this model:

a. The *'s represent parameters estimated using maximum likelihood, where constraints are allowed across structural components.

b. The solid lines, $\vec{b}$, relate the variance of the measured variables to its respective latent variable, *F1* or *F2*. Latent variables are unmeasured by definition with no relative location or scale, complicated somewhat by assuming distributional normality. Because normality is assumed and its first two moments are independent, each distributional parameter must be "fixed" separately. Setting the relationship, often referred to as a loading, between X3 and its appropriate factor to 1 fixes each factor's variance. Researchers commonly constrain the loadings of the same measured indicator in both structural components.

c. The dashed lines, $\vec{a}$, relate the mean of the measured variables to the latent variables, *F1* or *F2*. To simplify these relations, loadings between the measured variables and the intercept term, represented as a triangle in Figure 68, are constrained to be the same for both structural components. Second, the interest is in differences between the factor means, not individual values. To facilitate assessing these differences while fixing location, the loading between the intercept and *F1* is set to 0. This fixes the *F1* location, $\mu_{F1} = 0$, creating a baseline to compare the *F2* location, μ_{F2}.

Similar hypotheses from Chapter 1 are simultaneously considered; existence of treatment nonresponse, now at the latent level, differences in the factor means, and

equality of factor variances. The interest in latent means necessitates analysis of

augmented moment matrices, which take the form

$$
\mathbf{S}_A^2 = \begin{bmatrix} \ddots & \cdot^{\cdot} & \vdots \\ & \mathbf{S}^2 & \vec{\bar{\mathbf{x}}} \\ \cdot^{\cdot} & \ddots & \vdots \\ \cdots \vec{\mathbf{x}}' \cdots & 1 \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} & s_{13} & \bar{x}_1 \\ s_{21} & s_{22} & s_{23} & \bar{x}_2 \\ s_{31} & s_{32} & s_{33} & \bar{x}_3 \\ \bar{x}_1 & \bar{x}_2 & \bar{x}_3 & 1 \end{bmatrix}
\tag{138}
$$

for the augmented sample structure and

$$
\mathbf{\Sigma}_A = \begin{bmatrix} \ddots & \cdot^{\cdot} & \vdots \\ & \mathbf{\Sigma} & \vec{\bar{\mathbf{\mu}}} \\ \cdot^{\cdot} & \ddots & \vdots \\ \cdots \vec{\mathbf{\mu}}' \cdots & 1 \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \mu_1 \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \mu_2 \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \mu_3 \\ \mu_1 & \mu_2 & \mu_3 & 1 \end{bmatrix}
\tag{139}
$$

representing the model implied augmented moment structure, which are more

conveniently represented as a series of structural equations. These structural equations

contain the "*" values, the $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$ loadings, subsequently estimated when replacing

elements within the model implied augmented matrix. Maximum likelihood is still used

as the mechanism for parameter estimation, where first

$$
F = \frac{1}{2} \operatorname{tr} \left[([\mathbf{S}_A - \mathbf{\Sigma}_A(\hat{\mathbf{\theta}})] \, \mathbf{\Sigma}_A(\hat{\mathbf{\theta}})^{-1})^2 \right]
\tag{140}
$$

is calculated to provide the general form of the likelihood function for a single structural

component model as

$$
L(\mathbf{\theta}) = nF
\tag{141}
$$

where *tr* is the trace of the $r+1$ by $r+1$ augmented matrix consisting of the sample

augmented moment matrix $\mathbf{S}_A$ and model implied augmented moment matrix $\mathbf{\Sigma}_A(\hat{\mathbf{\theta}})$.

While the same *AIC* comparison can be used, models which posit treatment nonresponse cannot utilize summary representations in an augmented form for parameter estimation. Bollen (1989) provides the general form of the likelihood function for each individual in a single sample

$$L(\mathbf{\theta}) = \prod_{i=1}^{n} (2\pi)^{\frac{-p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\vec{\mathbf{x}}_i - \vec{\mathbf{\mu}})'\Sigma(\vec{\mathbf{x}}_i - \vec{\mathbf{\mu}})\right] \tag{142}$$

where *p* represent the number of parameters to be estimated for a particular model and the dimension of each vector equals the number of measured indicators, *r*. A natural log transformation can be performed, and multiplying each side by -2 results in

$$-2\ln(L(\mathbf{\theta})) = np\ln(2\pi) + n\ln|\Sigma| + \sum_{i=1}^{n}(\vec{\mathbf{x}}_i - \vec{\mathbf{\mu}})'\Sigma(\vec{\mathbf{x}}_i - \vec{\mathbf{\mu}}) \tag{143}$$

which is an *AIC* component. For a single sample, using similar notation from Figure 68, structural relationships exist where

$$\vec{\mathbf{\mu}} = \vec{\mathbf{x}} + \vec{\mathbf{a}}\mu_F \tag{144}$$

and

$$\Sigma = \vec{\mathbf{b}}\sigma_F^2\,\vec{\mathbf{b}}' + \mathbf{\Theta} \tag{145}$$

where Θ represents the dimension *r* covariance matrix of the errors of the form

$$\mathbf{\Theta} = \begin{bmatrix} \sigma_{e_1 e_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_{e_r e_r} \end{bmatrix} \tag{146}$$

with zeros in the off diagonals assuming their independence. These equations provide the framework to develop the likelihood function for two sample designs with treatment nonresponse on the latent variable.

Under a multivariate scenario, the K-S tests utilized as a means of defending the parametric distributions are no longer available. Annotated in Equation (134), each population is assumed to follow a multivariate normal distribution. Visualizing or even describing a parametric shape in more than two dimensions is challenging, where Mardia (1970, 1980) has advocated a series of omnibus tests evaluating this condition. Other less notable tests are also available (Mudholkar, 1992; Doornik, 1994). All of these methods evaluate both skewness and kurtosis which involve 3rd and 4th order moments in various combinations in their calculations.

With Mardia tests recommended as the multivariate replacement to the K-S tests, calculation of the test statistics requires a few steps. Considering first the control sample, the data is first centered where

$$\vec{\mathbf{x}}_c^c = (\vec{\mathbf{x}}_i - \overline{\mathbf{x}}_i) \quad \text{for } i = 1,..,r \tag{147}$$

and squared by

$$\mathbf{M}_{mat} = \frac{\vec{\mathbf{x}}_c^{c\,\prime} \vec{\mathbf{x}}_c^c}{n} \tag{148}$$

resulting in a r by r matrix. This matrix is then transformed to a n_C by n_C matrix whose result is used in the calculation of the Mardia skewness and kurtosis test statistics by

$$\mathbf{R}_{ij} = \vec{\mathbf{x}}_c^c \, \mathbf{M}_{mat} \, \vec{\mathbf{x}}_c^{c\,\prime} \tag{149}$$

Calculation of the skewness test statistic is

$$M_{sk} = \frac{n \left(\dfrac{1}{n^2} \displaystyle\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{R}_{ij}^3 \right)}{6} \tag{150}$$

which asymptotically follows a chi-square distribution whose degrees of freedom equals

151

$$df = \frac{r(r+1)(r+2)}{6} \tag{151}$$

The kurtosis test statistic is calculated

$$M_{\text{kurt}} = \frac{\left(\dfrac{\text{tr}(\mathbf{R}_{ij}^{2})}{n}\right) - r(r+2)}{\sqrt{\dfrac{8(r)(r+2)}{n}}} \tag{152}$$

which asymptotically subscribes to a standard normal distribution. These test statistics are the basis evaluating the following hypothesis

$Ho: F(\vec{x}_C) = F^*(\vec{x}_C)$ - the data subscribes to a multivariate normal distribution

$Ha: F(\vec{x}_C) \neq F^*(\vec{x}_C)$ - the data does not subscribe to a multivariate normal distribution

where the null hypothesis would be rejected if either test statistic exceeded its respective critical value against a specified Type I error control.

Evaluating a similar hypothesis with the treatment sample requires an additional step where, unlike K-S tests, Mardia's tests are not inherently supportive of mixtures. Assessing multivariate normality of the treatment sample can be conducted by weighting each individual's set of responses according to the complement of their posterior probability of being classified as a treatment nonrespondent represented as

$$\omega_j = (1 - \hat{\pi}_j) \tag{153}$$

determined post min *AIC* selection using Bayes' theorem. Weighting cases or respondents is common in surveys to account for oversampling, but applied in this manner separates any nonresponse impact from the sample establishing pseudo *ID* conditions. Any model selection failing to support treatment nonresponse results in each respondent's weighting as 1. The construction and evaluation of the multivariate

normality hypothesis for the weighted treatment sample proceeds in the same manner outlined above.

4.4 Closing Remarks

The development and proliferation of computers transitioned finite mixture models from theoretical development to practical application by handling the complex process of parameter estimation. More recently, significant advances in computing power and customizable programming packages have enabled comprehensive empirical studies in justifying their utility and applicability across many fields of research. Consequently, the collective result has been an increased popularity for this *newer* type of modeling where its integration with or replacement of traditional techniques has only begun to be developed.

Similarly, the use of information criterion methods and assessment of multiple representations of a data structure is much more available now than in recent years. Assessing multiple representations in an omnibus fashion provides many advantages over probabilistic inferences on single parameter estimates burdened with normality requirements and the strong assumption of a correct model specification. As shown, information criterion measures can be used as an alternative to basic population means and variance hypothesis tests, assessed simultaneously, without requiring multiple comparison corrections, or conditioning results on a subjectively chosen Type I error control. For those uneasy about subscribing to a data driven model selection strategy, this is the same data providing empirical evidence for the inferential testing of multiple model parameter estimates.

As research questions become more involved, so does the corresponding complexity of models required to provide empirical evidence. It should naturally follow, therefore, to obtain evidence in an omnibus sense rather than dissect complex models to a series of independent hypothesis tests of parameter estimates that are invariably related. When considering finite mixture models, such dissection can not be done, leaving an omnibus assessment as the only available technique to obtain empirical evidence. Finally, and most exciting, are the numerous extensions available from the basic methodological construct advocated in this dissertation, particularly the multivariate extension.

Appendix 1: Normal Distribution Experimental Study Conditions

VARIANCE RATIO = 1.0

Empirical Conditions for $\sigma^2_{rat} = 1.00$, $n_C = n_T$ at $\phi = 0.00$

z^*	$n_T = 50$		$n_T = 100$		$n_T = 200$		$n_T = 350$	
	d^*	σ_C, σ_T	d^*	σ_C, σ_T	d^*	σ_C, σ_T	d^*	σ_C, σ_T
3	0.60	16.65	0.42	23.60	0.30	33.30	$d^* < 0.25$	
4	0.80	12.49	0.57	17.66	0.40	25.00	0.30	33.10
7	1.40	7.14	0.99	10.10	0.70	14.30	0.53	18.89
10	2.00	5.00	1.41	7.07	1.00	10.00	0.76	13.23
15	3.00	3.33	2.12	4.71	1.50	6.66	1.13	8.82
20	4.00	2.50	2.82	3.54	2.00	5.00	1.51	6.61
25	$d^* > 4$		3.53	2.83	2.50	4.00	1.89	5.29
30	$d^* > 4$		$d^* > 4$		3.00	3.33	2.27	4.41
40	$d^* > 4$		$d^* > 4$		4.00	2.50	3.02	3.31

Empirical Conditions for $\sigma^2_{rat} = 1.00$, $n_C = n_T$ at $\phi = 0.05$

z^*	$n_T = 50$		$n_T = 100$		$n_T = 200$		$n_T = 350$	
	d^*	σ_C, σ_T	d^*	σ_C, σ_T	d^*	σ_C, σ_T	d^*	σ_C, σ_T
3	0.60	16.65	0.42	23.55	0.30	33.30	$d^* < 0.25$	
4	0.80	12.47	0.57	17.65	0.40	24.95	0.30	33.00
7	1.40	7.13	0.99	10.09	0.70	14.30	0.53	18.88
10	2.00	4.99	1.42	7.06	1.00	10.00	0.76	13.21
15	3.00	3.33	2.12	4.71	1.50	6.66	1.14	8.81
20	4.00	2.50	2.83	3.53	2.00	4.99	1.51	6.61
25	$d^* > 4$		3.55	2.82	2.50	4.00	1.89	5.28
30	$d^* > 4$		$d^* > 4$		3.00	3.33	2.27	4.40
40	$d^* > 4$		$d^* > 4$		4.00	2.50	3.03	3.30

Empirical Conditions for $\sigma^2_{rat} = 1.00$, $n_C = n_T$ at $\phi = 0.10$

z^*	$n_T = 50$		$n_T = 100$		$n_T = 200$		$n_T = 350$	
	d^*	σ_C, σ_T	d^*	σ_C, σ_T	d^*	σ_C, σ_T	d^*	σ_C, σ_T
3	0.60	16.60	0.42	23.60	0.30	33.20	$d^* < 0.25$	
4	0.80	12.45	0.57	17.60	0.40	24.90	0.30	32.90
7	1.41	7.11	1.00	10.05	0.70	14.22	0.53	18.80
10	2.01	4.97	1.42	7.03	1.01	9.95	0.76	13.16
15	3.01	3.32	2.13	4.69	1.51	6.63	1.14	8.77
20	4.00	2.49	2.84	3.52	2.01	4.97	1.52	6.58
25	$d^* > 4$		3.56	2.81	2.51	3.98	1.90	5.26
30	$d^* > 4$		$d^* > 4$		3.01	3.32	2.28	4.39
40	$d^* > 4$		$d^* > 4$		4.00	2.49	3.04	3.29

Appendix 1: Normal Distribution Experimental Study Conditions (continued)

Empirical Conditions for $\sigma^2_{rat} = 1.00$, $n_C = n_T$ at $\phi = 0.20$

z^*	$n_T = 50$		$n_T = 100$		$n_T = 200$		$n_T = 350$	
	d^*	$\sigma_{C,}\ \sigma_T$	d^*	$\sigma_{C,}\ \sigma_T$	d^*	$\sigma_{C,}\ \sigma_T$	d^*	$\sigma_{C,}\ \sigma_T$
3	0.61	16.35	0.43	23.10	0.31	32.70	$d^* < 0.25$	
4	0.82	12.25	0.58	17.30	0.41	24.50	0.31	32.40
7	1.43	7.00	1.01	9.90	0.71	14.00	0.54	18.51
10	2.04	4.90	1.44	6.93	1.02	9.80	0.77	12.96
15	3.06	3.27	2.16	4.62	1.53	6.53	1.16	8.64
20	$d^* > 4$		2.89	3.46	2.04	4.90	1.54	6.48
25	$d^* > 4$		3.61	2.77	2.55	3.92	1.93	5.18
30	$d^* > 4$		$d^* > 4$		3.06	3.27	2.31	4.32
40	$d^* > 4$		$d^* > 4$		$d^* > 4$		3.09	3.24

Empirical Conditions for $\sigma^2_{rat} = 1.00$, $n_C = n_T$ at $\phi = 0.35$

z^*	$n_T = 50$		$n_T = 100$		$n_T = 200$		$n_T = 350$	
	d^*	$\sigma_{C,}\ \sigma_T$	d^*	$\sigma_{C,}\ \sigma_T$	d^*	$\sigma_{C,}\ \sigma_T$	d^*	$\sigma_{C,}\ \sigma_T$
3	0.64	15.55	0.45	22.10	0.32	31.20	$d^* < 0.25$	
4	0.86	11.67	0.60	16.55	0.43	23.40	0.32	31.00
7	1.50	6.66	1.06	9.45	0.75	13.39	0.56	17.70
10	2.14	4.67	1.51	6.63	1.07	9.37	0.81	12.39
15	3.22	3.11	2.26	4.42	1.60	6.25	1.22	8.20
20	$d^* > 4$		3.02	3.31	2.14	4.68	1.61	6.20
25	$d^* > 4$		3.77	2.65	2.67	3.75	2.02	4.95
30	$d^* > 4$		$d^* > 4$		3.21	3.12	2.42	4.13
40	$d^* > 4$		$d^* > 4$		$d^* > 4$		3.23	3.10

Empirical Conditions for $\sigma^2_{rat} = 1.00$, $n_C = n_T$ at $\phi = 0.50$

z^*	$n_T = 50$		$n_T = 100$		$n_T = 200$		$n_T = 350$	
	d^*	$\sigma_{C,}\ \sigma_T$	d^*	$\sigma_{C,}\ \sigma_T$	d^*	$\sigma_{C,}\ \sigma_T$	d^*	$\sigma_{C,}\ \sigma_T$
3	0.69	14.45	0.49	20.40	0.35	28.90	0.26	38.20
4	0.92	10.83	0.65	15.30	0.46	21.65	0.35	28.65
7	1.62	6.19	1.14	8.75	0.81	12.37	0.61	16.37
10	2.31	4.33	1.63	6.13	1.15	8.66	0.87	11.45
15	3.46	2.89	2.45	4.08	1.73	5.77	1.31	7.65
20	$d^* > 4$		3.27	3.06	2.31	4.33	1.75	5.73
25	$d^* > 4$		$d^* > 4$		2.89	3.46	2.18	4.58
30	$d^* > 4$		$d^* > 4$		3.46	2.89	2.62	3.82
40	$d^* > 4$		$d^* > 4$		$d^* > 4$		3.50	2.86

Appendix 1: Normal Distribution Experimental Study Conditions (continued)

VARIANCE RATIO = 0.50

Empirical Conditions for $\sigma^2_{rat} = 0.50$, $n_C = n_T$ at $\phi = 0.00$

	$n_T = 100$			$n_T = 200$			$n_T = 350$		
z^*	d^*	σ_C	σ_T	d^*	σ_C	σ_T	d^*	σ_C	σ_T
3	0.42	19.25	27.22	0.30	27.25	38.54		$d^* < 0.25$	
4	0.57	14.42	20.39	0.40	20.40	28.85	0.30	27.00	38.18
7	1.00	8.25	11.67	0.71	11.66	16.49	0.54	15.43	21.82
10	1.44	5.77	8.16	1.01	8.17	11.55	0.77	10.80	15.27
15	2.15	3.85	5.44	1.52	5.44	7.69	1.15	7.20	10.18
20	2.87	2.89	4.09	2.03	4.08	5.77	1.53	5.40	7.64
25	3.59	2.31	3.27	2.53	3.27	4.62	1.92	4.32	6.11
30		$d^* > 4$		3.05	2.72	3.85	2.30	3.60	5.09
40		$d^* > 4$			$d^* > 4$		3.07	2.70	3.82

Empirical Conditions for $\sigma^2_{rat} = 0.50$, $n_C = n_T$ at $\phi = 0.05$

	$n_T = 100$			$n_T = 200$			$n_T = 350$		
z^*	d^*	σ_C	σ_T	d^*	σ_C	σ_T	d^*	σ_C	σ_T
3	0.43	19.05	26.94	0.30	27.00	38.18		$d^* < 0.25$	
4	0.58	14.30	20.22	0.41	20.20	28.57	0.31	26.75	37.83
7	1.02	8.17	11.55	0.72	11.55	16.33	0.55	15.29	21.62
10	1.46	5.72	8.09	1.03	8.09	11.44	0.78	10.70	15.13
15	2.19	3.81	5.39	1.55	5.39	7.62	1.17	7.13	10.08
20	2.92	2.86	4.04	2.07	4.04	5.71	1.56	5.35	7.57
25	3.65	2.29	3.24	2.59	3.23	4.57	1.95	4.28	6.05
30		$d^* > 4$		3.09	2.70	3.82	2.35	3.56	5.03
40		$d^* > 4$			$d^* > 4$		3.13	2.67	3.78

Empirical Conditions for $\sigma^2_{rat} = 0.50$, $n_C = n_T$ at $\phi = 0.10$

	$n_T = 100$			$n_T = 200$			$n_T = 350$		
z^*	d^*	σ_C	σ_T	d^*	σ_C	σ_T	d^*	σ_C	σ_T
3	0.44	18.85	26.66	0.31	26.60	37.62		$d^* < 0.25$	
4	0.60	14.13	19.98	0.42	20.00	28.28	0.31	26.40	37.34
7	1.04	8.07	11.41	0.74	11.41	16.14	0.56	15.10	21.35
10	1.49	5.65	7.99	1.05	7.99	11.30	0.80	10.57	14.95
15	2.24	3.77	5.33	1.58	5.33	7.54	1.20	7.05	9.97
20	2.98	2.83	4.00	2.11	4.00	5.66	1.59	5.29	7.48
25	3.73	2.26	3.20	2.63	3.20	4.53	1.99	4.23	5.98
30		$d^* > 4$		3.17	2.66	3.76	2.39	3.52	4.98
40		$d^* > 4$			$d^* > 4$		3.19	2.64	3.73

Appendix 1: Normal Distribution Experimental Study Conditions (continued)

Empirical Conditions for $\sigma^2_{rat} = 0.50$, $n_C = n_T$ at $\phi = 0.20$

z^*	$n_T = 100$			$n_T = 200$			$n_T = 350$		
	d^*	σ_C	σ_T	d^*	σ_C	σ_T	d^*	σ_C	σ_T
3	0.46	18.25	25.81	0.33	25.80	36.49	$d^* < 0.25$		
4	0.63	13.70	19.37	0.44	19.38	27.41	0.33	25.60	36.20
7	1.10	7.82	11.06	0.77	11.07	15.66	0.59	14.65	20.72
10	1.57	5.47	7.74	1.11	7.75	10.96	0.84	10.25	14.50
15	2.35	3.65	5.16	1.66	5.16	7.30	1.26	6.83	9.66
20	3.13	2.74	3.87	2.22	3.87	5.47	1.68	5.12	7.24
25	3.92	2.19	3.10	2.77	3.10	4.38	2.09	4.10	5.80
30	$d^* > 4$			3.33	2.58	3.65	2.51	3.42	4.84
40	$d^* > 4$			$d^* > 4$			3.35	2.56	3.62

Empirical Conditions for $\sigma^2_{rat} = 0.50$, $n_C = n_T$ at $\phi = 0.35$

z^*	$n_T = 100$			$n_T = 200$			$n_T = 350$		
	d^*	σ_C	σ_T	d^*	σ_C	σ_T	d^*	σ_C	σ_T
3	0.51	17.05	24.11	0.36	24.10	34.08	$d^* < 0.25$		
4	0.69	12.80	18.10	0.48	18.10	25.60	0.36	23.90	33.80
7	1.21	7.31	10.34	0.85	10.34	14.62	0.64	13.67	19.33
10	1.72	5.12	7.24	1.22	7.24	10.24	0.92	9.57	13.53
15	2.58	3.41	4.82	1.82	4.83	6.83	1.38	6.38	9.02
20	3.44	2.56	3.62	2.43	3.62	5.12	1.84	4.78	6.76
25	$d^* > 4$			3.05	2.89	4.09	2.30	3.83	5.42
30	$d^* > 4$			3.66	2.41	3.41	2.76	3.19	4.51
40	$d^* > 4$			$d^* > 4$			3.69	2.39	3.38

Empirical Conditions for $\sigma^2_{rat} = 0.50$, $n_C = n_T$ at $\phi = 0.50$

z^*	$n_T = 100$			$n_T = 200$			$n_T = 350$		
	d^*	σ_C	σ_T	d^*	σ_C	σ_T	d^*	σ_C	σ_T
3	0.58	15.45	21.85	0.41	21.80	30.83	$d^* < 0.25$		
4	0.78	11.58	16.38	0.55	16.35	23.12	0.41	21.65	30.62
7	1.37	6.61	9.35	0.97	9.35	13.22	0.73	12.38	17.51
10	1.96	4.63	6.55	1.38	6.55	9.26	1.05	8.66	12.25
15	2.93	3.09	4.37	2.08	4.36	6.17	1.57	5.77	8.16
20	3.92	2.31	3.27	2.77	3.27	4.62	2.09	4.33	6.12
25	$d^* > 4$			3.46	2.62	3.71	2.62	3.46	4.89
30	$d^* > 4$			$d^* > 4$			3.14	2.89	4.09
40	$d^* > 4$			$d^* > 4$			$d^* > 4$		

Appendix 1: Normal Distribution Experimental Study Conditions (continued)

VARIANCE RATIO = 2.0

Empirical Conditions for $\sigma^2_{rat} = 2.00$, $n_C = n_T$ at $\phi = 0.00$

	$n_T = 100$			$n_T = 200$			$n_T = 350$		
z^*	d^*	σ_C	σ_T	d^*	σ_C	σ_T	d^*	σ_C	σ_T
3	0.42	27.22	19.25	0.30	38.54	27.25		$d^* < 0.25$	
4	0.57	20.39	14.42	0.40	28.85	20.40	0.30	38.18	27.00
7	1.00	11.67	8.25	0.71	16.49	11.66	0.54	21.82	15.43
10	1.44	8.16	5.77	1.01	11.55	8.17	0.77	15.27	10.80
15	2.15	5.44	3.85	1.52	7.69	5.44	1.15	10.18	7.20
20	2.87	4.09	2.89	2.03	5.77	4.08	1.53	7.64	5.40
25	3.59	3.27	2.31	2.53	4.62	3.27	1.92	6.11	4.32
30		$d^* > 4$		3.05	3.85	2.72	2.30	5.09	3.60
40		$d^* > 4$			$d^* > 4$		3.07	3.82	2.70

Empirical Conditions for $\sigma^2_{rat} = 2.00$, $n_C = n_T$ at $\phi = 0.05$

	$n_T = 100$			$n_T = 200$			$n_T = 350$		
z^*	d^*	σ_C	σ_T	d^*	σ_C	σ_T	d^*	σ_C	σ_T
3	0.42	27.40	19.37	0.30	38.80	27.44		$d^* < 0.25$	
4	0.57	20.55	14.53	0.39	29.10	20.58	0.30	38.50	27.22
7	0.99	11.75	8.31	0.70	16.61	11.75	0.53	21.99	15.55
10	1.41	8.22	5.81	1.00	11.63	8.22	0.75	15.39	10.88
15	2.12	5.48	3.87	1.50	7.76	5.49	1.13	10.25	7.25
20	2.83	4.11	2.91	2.00	5.82	4.12	1.51	7.69	5.44
25	3.53	3.29	2.33	2.50	4.65	3.29	1.89	6.15	4.35
30		$d^* > 4$		2.99	3.88	2.74	2.26	5.13	3.63
40		$d^* > 4$		3.99	2.91	2.06	3.02	3.85	2.72

Empirical Conditions for $\sigma^2_{rat} = 2.00$, $n_C = n_T$ at $\phi = 0.10$

	$n_T = 100$			$n_T = 200$			$n_T = 350$		
z^*	d^*	σ_C	σ_T	d^*	σ_C	σ_T	d^*	σ_C	σ_T
3	0.41	27.50	19.45	0.29	39.00	27.58		$d^* < 0.25$	
4	0.56	20.65	14.60	0.39	29.20	20.65	0.29	38.60	27.29
7	0.98	11.80	8.34	0.69	16.70	11.81	0.52	22.09	15.62
10	1.39	8.26	5.84	0.99	11.69	8.27	0.75	15.46	10.93
15	2.09	5.51	3.90	1.48	7.79	5.51	1.12	10.31	7.29
20	2.79	4.13	2.92	1.97	5.84	4.13	1.49	7.73	5.47
25	3.48	3.31	2.34	2.46	4.68	3.31	1.86	6.18	4.37
30		$d^* > 4$		2.96	3.89	2.75	2.24	5.15	3.64
40		$d^* > 4$		3.94	2.92	2.06	2.98	3.87	2.74

Appendix 1: Normal Distribution Experimental Study Conditions (continued)

Empirical Conditions for $\sigma^2_{rat} = 2.00$, $n_C = n_T$ at $\phi = 0.20$

	$n_T = 100$			$n_T = 200$			$n_T = 350$		
z^*	d^*	σ_C	σ_T	d^*	σ_C	σ_T	d^*	σ_C	σ_T
3	0.41	27.60	19.52	0.29	39.00	27.58		$d^* < 0.25$	
4	0.55	20.70	14.64	0.38	29.30	20.72	0.29	38.70	27.37
7	0.96	11.83	8.37	0.68	16.72	11.82	0.51	22.13	15.65
10	1.37	8.28	5.85	0.97	11.71	8.28	0.73	15.49	10.95
15	2.05	5.52	3.90	1.45	7.81	5.52	1.10	10.33	7.30
20	2.74	4.14	2.93	1.93	5.86	4.14	1.46	7.75	5.48
25	3.42	3.31	2.34	2.42	4.69	3.32	1.83	6.20	4.38
30		$d^* > 4$		2.90	3.90	2.76	2.19	5.17	3.66
40		$d^* > 4$		3.87	2.93	2.07	2.93	3.87	2.74

Empirical Conditions for $\sigma^2_{rat} = 2.00$, $n_C = n_T$ at $\phi = 0.35$

	$n_T = 100$			$n_T = 200$			$n_T = 350$		
z^*	d^*	σ_C	σ_T	d^*	σ_C	σ_T	d^*	σ_C	σ_T
3	0.40	27.10	19.16	0.29	38.30	27.08		$d^* < 0.25$	
4	0.54	20.35	14.39	0.38	28.80	20.36	0.29	38.05	26.91
7	0.95	11.63	8.22	0.67	16.45	11.63	0.51	21.75	15.38
10	1.36	8.14	5.76	0.96	11.51	8.14	0.73	15.22	10.76
15	2.04	5.43	3.84	1.44	7.67	5.42	1.09	10.15	7.18
20	2.72	4.07	2.88	1.92	5.76	4.07	1.45	7.61	5.38
25	3.39	3.26	2.31	2.40	4.60	3.25	1.81	6.09	4.31
30		$d^* > 4$		2.88	3.84	2.72	2.18	5.07	3.59
40		$d^* > 4$		3.84	2.88	2.04	2.90	3.81	2.69

Empirical Conditions for $\sigma^2_{rat} = 2.00$, $n_C = n_T$ at $\phi = 0.50$

	$n_T = 100$			$n_T = 200$			$n_T = 350$		
z^*	d^*	σ_C	σ_T	d^*	σ_C	σ_T	d^*	σ_C	σ_T
3	0.41	25.80	18.24	0.29	36.50	25.81		$d^* < 0.25$	
4	0.56	19.35	13.68	0.39	27.40	19.37	0.30	36.20	25.60
7	0.97	11.07	7.83	0.69	15.65	11.07	0.52	20.70	14.64
10	1.39	7.74	5.47	0.99	10.95	7.74	0.74	14.49	10.25
15	2.09	5.17	3.66	1.48	7.31	5.17	1.12	9.66	6.83
20	2.79	3.87	2.74	1.97	5.48	3.87	1.49	7.25	5.13
25	3.48	3.10	2.19	2.46	4.38	3.10	1.86	5.80	4.10
30		$d^* > 4$		2.96	3.65	2.58	2.23	4.83	3.42
40		$d^* > 4$		3.94	2.74	1.94	2.98	3.62	2.56

Appendix 1: Normal Distribution Experimental Study Conditions (continued)

VARIANCE RATIO = 0.33

Empirical Conditions for $\sigma^2_{rat} = 0.33$, $n_C = n_T$ at $n_T = 100$

z^*	$\phi = 0.0$			$\phi = 0.10$			$\phi = 0.20$			$\phi = 0.35$		
	d^*	σ_C	σ_T	d^*	σ_C	σ_T	d^*	σ_C	σ_T	d^*	σ_C	σ_T
3	0.42	16.65	28.84	0.45	16.20	28.06	0.48	15.55	26.93	0.54	14.40	24.94
4	0.57	12.50	21.65	0.60	12.15	21.04	0.64	11.67	20.21	0.72	10.80	18.71
7	0.99	7.14	12.37	1.05	6.94	12.02	1.12	6.67	11.55	1.26	6.18	10.70
10	1.41	5.00	8.66	1.50	4.85	8.40	1.60	4.67	8.09	1.80	4.32	7.48
15	2.12	3.33	5.77	2.24	3.24	5.61	2.40	3.11	5.39	2.70	2.88	4.99
20	2.83	2.50	4.33	2.99	2.43	4.21	3.19	2.34	4.05	3.60	2.16	3.74
25	3.54	2.00	3.46	3.74	1.94	3.36	$d^* > 4$			$d^* > 4$		
30	$d^* > 4$			$d^* > 4$			$d^* > 4$			$d^* > 4$		
40	$d^* > 4$			$d^* > 4$			$d^* > 4$			$d^* > 4$		

Empirical Conditions for $\sigma^2_{rat} = 0.33$, $n_C = n_T$ at $n_T = 200$

z^*	$\phi = 0.0$			$\phi = 0.10$			$\phi = 0.20$			$\phi = 0.35$		
	d^*	σ_C	σ_T	d^*	σ_C	σ_T	d^*	σ_C	σ_T	d^*	σ_C	σ_T
3	0.30	23.60	40.88	0.32	22.90	39.66	0.34	22.00	38.11	0.38	20.40	35.33
4	0.40	17.68	30.62	0.42	17.15	29.70	0.45	16.50	28.58	0.51	15.28	26.47
7	0.70	10.10	17.49	0.74	9.81	16.99	0.79	9.44	16.35	0.89	8.73	15.12
10	1.00	7.07	12.25	1.06	6.86	11.88	1.13	6.60	11.43	1.27	6.11	10.58
15	1.50	4.71	8.16	1.58	4.58	7.93	1.69	4.40	7.62	1.91	4.07	7.05
20	2.00	3.53	6.11	2.12	3.43	5.94	2.26	3.30	5.72	2.54	3.06	5.30
25	2.50	2.83	4.90	2.64	2.75	4.76	2.82	2.64	4.57	3.19	2.44	4.23
30	3.00	2.36	4.09	3.17	2.29	3.97	3.39	2.20	3.81	3.82	2.04	3.53
40	$d^* > 4$			$d^* > 4$			$d^* > 4$			$d^* > 4$		

Appendix 1: Normal Distribution Experimental Study Conditions (continued)

VARIANCE RATIO = 3.0

Empirical Conditions for $\sigma^2_{rat} = 3.00$, $n_C = n_T$ at $n_T = 100$

z^*	$\phi = 0.0$			$\phi = 0.10$			$\phi = 0.20$			$\phi = 0.35$		
	d^*	σ_C	σ_T	d^*	σ_C	σ_T	d^*	σ_C	σ_T	d^*	σ_C	σ_T
3	0.42	28.84	16.65	0.41	29.50	17.03	0.39	29.80	17.21	0.38	29.80	17.21
4	0.57	21.65	12.50	0.54	22.10	12.76	0.52	22.38	12.92	0.51	22.35	12.90
7	0.99	12.37	7.14	0.95	12.63	7.29	0.91	12.78	7.38	0.89	12.76	7.37
10	1.41	8.66	5.00	1.35	8.84	5.10	1.31	8.94	5.16	1.27	8.93	5.16
15	2.12	5.77	3.33	2.03	5.89	3.40	1.96	5.96	3.44	1.90	5.95	3.44
20	2.83	4.33	2.50	2.70	4.42	2.55	2.61	4.47	2.58	2.53	4.46	2.57
25	3.54	3.46	2.00	3.38	3.54	2.04	3.26	3.58	2.07	3.16	3.57	2.06
30	$d^* > 4$			$d^* > 4$			3.92	2.98	1.72	3.79	2.98	1.72
40	$d^* > 4$			$d^* > 4$			$d^* > 4$			$d^* > 4$		

Empirical Conditions for $\sigma^2_{rat} = 3.00$, $n_C = n_T$ at $n_T = 200$

z^*	$\phi = 0.0$			$\phi = 0.10$			$\phi = 0.20$			$\phi = 0.35$		
	d^*	σ_C	σ_T	d^*	σ_C	σ_T	d^*	σ_C	σ_T	d^*	σ_C	σ_T
3	0.30	40.88	23.60	0.29	41.70	24.08	0.28	42.10	24.31	0.27	42.10	24.31
4	0.40	30.62	17.68	0.38	31.25	18.04	0.37	31.60	18.24	0.36	31.60	18.24
7	0.70	17.50	10.10	0.67	17.85	10.31	0.65	18.08	10.44	0.63	18.05	10.42
10	1.00	12.25	7.07	0.96	12.50	7.22	0.92	12.65	7.30	0.89	12.63	7.29
15	1.50	8.17	4.72	1.43	8.34	4.82	1.39	8.43	4.87	1.34	8.42	4.86
20	2.00	6.12	3.53	1.91	6.25	3.61	1.85	6.32	3.65	1.79	6.32	3.65
25	2.50	4.90	2.83	2.39	5.00	2.89	2.31	5.06	2.92	2.24	5.05	2.92
30	3.00	4.08	2.36	2.87	4.17	2.41	2.77	4.22	2.44	2.68	4.21	2.43
40	4.00	3.06	1.77	3.82	3.13	1.81	3.70	3.16	1.82	3.58	3.16	1.82

Appendix 2: Correct Treatment Effect Hypothesis Conclusion %

| | | $z^* = 3$ | | | | $z^* = 4$ | | | | $z^* = 7$ | | | |
| | | n_T | | | | n_T | | | | n_T | | | |
σ^2_{rat}	ϕ	50	100	200	350	50	100	200	350	50	100	200	350
0.33	0		95.6	96.2			100.0	100.0			100.0	100.0	
	10		92.2	93.2			98.6	98.8			100.0	100.0	
	20		90.8	94.4			99.6	99.6			100.0	100.0	
	35		92.2	90.4			98.4	98.6			100.0	100.0	
0.50	0		95.2	93.8			99.6	100.0	99.4		100.0	100.0	100.0
	5		96.0	92.2			99.4	98.8	99.6		100.0	100.0	100.0
	10		93.0	91.4			99.6	99.2	99.8		100.0	100.0	100.0
	20		89.8	92.4			96.6	98.4	97.4		100.0	100.0	100.0
	35		85.0	85.8			96.4	96.6	95.8		100.0	100.0	100.0
	50		78.8	81.0			92.4	92.0	94.0		100.0	100.0	100.0
1.00	0	94.4	95.0	94.4			99.4	99.0	99.8	100.0	100.0	100.0	100.0
	5	92.4	93.6	92.0			98.8	99.6	99.0	100.0	100.0	100.0	100.0
	10	90.8	88.4	94.0			98.8	99.0	98.4	100.0	100.0	100.0	100.0
	20	86.4	89.2	85.2			97.8	95.4	96.6	100.0	100.0	100.0	100.0
	35	76.8	76.6	74.8			90.0	91.4	90.6	100.0	100.0	99.8	99.8
	50	60.2	65.2	66.6	61.2		81.4	81.8	81.4	98.8	99.2	100.0	99.6
2.00	0		94.2	93.2			99.8	99.8	99.8		100.0	100.0	100.0
	5		95.2	92.2			98.4	99.0	99.2		100.0	100.0	100.0
	10		93.4	89.0			99.0	98.0	99.0		100.0	100.0	100.0
	20		84.4	85.2			95.2	97.0	96.8		100.0	100.0	100.0
	35		76.6	76.8			91.2	90.6	90.6		99.8	100.0	100.0
	50		66.8	68.6			75.0	83.8	81.4		95.8	98.6	98.2
3.00	0		95.2	93.0			100.0	99.2			100.0	100.0	
	10		94.0	92.8			99.2	98.4			100.0	100.0	
	20		91.0	91.4			98.4	97.2			100.0	100.0	
	35		83.4	88.0			93.4	94.0			99.6	100.0	

Appendix 3: False Mixture Classifcation % when $\phi = 0.0$

Note: Entries in emboldened italics correspond to correct treatment effect mean hypothesis conclusions below 100%.

n_T	σ^2_{rat}	z^* 3	4	7	10	15	20	25	30	40
50	1.00	*10.2*	*9.8*	8.8	7.0	3.8	1.8			
100	0.33	*8.2*	8.2	21.8	15.4	9.4	6.6	3.8		
	0.50	*10.4*	*15.6*	32.8	18.8	7.6	5.8	4.0		
	1.00	*2.6*	*9.6*	11.0	9.2	7.2	6.4	2.2		
	2.00	*7.4*	*4.2*	5.6	4.2	3.8	3.2	2.0		
	3.00	*5.0*	3.4	4.8	3.8	3.8	1.4	0.8		
200	0.33	*7.4*	8.4	11.4	12.8	12.4	7.4	7.4	7.4	
	0.50	*5.8*	5.0	19.8	22.6	8.6	9.0	7.4	6.6	
	1.00	*1.6*	*4.6*	7.4	8.2	8.2	5.2	6.6	5.2	1.8
	2.00	*5.4*	*7.4*	5.6	4.8	5.0	2.2	4.0	2.4	
	3.00	*3.6*	*4.4*	4.2	3.2	4.2	4.0	2.2	1.6	0.6
350	0.50		*9.4*	8.4	21.4	14.8	10.2	8.6	7.6	7.6
	1.00		*2.6*	9.4	8.0	7.2	7.0	5.4	5.8	4.6
	2.00		*4.0*	4.2	5.2	5.0	3.2	3.6	3.4	2.0

Appendix 4: Correct Mixture Hypothesis Conclusion %

Note: Entries in emboldened italics correspond to correct treatment effect mean hypothesis conclusions below 100%.

$\sigma^2_{rat} = 0.33$

						z^*				
n_T	ϕ	3	4	7	10	15	20	25	30	40
100	10	*16.6*	*17.2*	31.4	30.0	36.8	71.4	97.2		
	20	*23.2*	*31.8*	45.6	52.6	66.8	96.4			
	35	*46.8*	*58.4*	76.0	86.4	96.2	100.0			
200	10	*15.6*	*15.6*	27.2	33.6	38.0	44.2	74.6	96.6	
	20	*34.2*	*37.2*	48.2	59.6	69.4	86.0	97.2	100.0	
	35	*55.8*	*62.6*	83.0	92.8	96.6	99.8	100.0	100.0	

$\sigma^2_{rat} = 0.50$

						z^*				
n_T	ϕ	3	4	7	10	15	20	25	30	40
100	5	*12.0*	*20.4*	34.0	22.2	22.0	57.0	90.0		
	10	*12.6*	*23.2*	38.6	24.8	32.2	79.2	99.0		
	20	*17.4*	*30.4*	48.4	37.8	56.0	94.8	99.8		
	35	*28.8*	*46.4*	59.8	63.8	86.4	98.8			
	50	*43.0*	*58.2*	79.4	87.2	99.2	100.0			
200	5	*8.8*	*10.6*	26.0	26.8	14.8	25.6	54.0	85.0	
	10	*11.2*	*15.0*	31.6	30.4	18.4	37.4	81.0	98.8	
	20	*15.8*	*16.6*	43.8	42.6	45.0	66.2	95.2	100.0	
	35	*28.0*	*38.4*	63.8	70.6	82.2	92.2	99.2	100.0	
	50	*37.2*	*53.2*	82.0	90.2	98.4	100.0	100.0		
350	5		*10.8*	16.2	20.6	21.6	21.8	24.8	50.8	98.0
	10		*11.0*	16.6	33.0	26.6	29.8	41.2	77.8	100.0
	20		*27.2*	36.2	50.4	51.4	59.4	77.8	96.2	100.0
	35		*39.2*	63.4	79.6	87.6	93.8	98.0	100.0	100.0
	50		*56.2*	84.0	95.6	99.2	99.8	100.0	100.0	

Appendix 4: Correct Mixture Hypothesis Conclusion %
(continued)

Note: Entries in emboldened italics correspond to correct treatment effect mean hypothesis conclusions below 100%.

$\sigma^2_{rat} = 1.0$

						z*				
n_T	ϕ	3	4	7	10	15	20	25	30	40
50	5	***10.4***	***13.6***	15.6	34.2	73.6	96.4			
	10	***10.0***	***13.0***	27.8	49.2	87.6	99.4			
	20	***11.6***	***16.6***	35.4	65.4	94.2				
	35	***13.6***	***20.4***	46.2	72.0	97.2				
	50	***11.8***	***23.2***	***49.8***	79.8	98.4				
100	5	***3.2***	***11.2***	13.8	21.2	48.6	80.0	96.4		
	10	***5.8***	***9.6***	20.0	33.8	68.2	96.0	100.0		
	20	***6.8***	***9.4***	21.6	52.0	83.2	98.6	100.0		
	35	***5.8***	***12.2***	31.0	56.8	88.0	99.8	99.6		
	50	***8.2***	***16.6***	***39.8***	64.8	90.8	99.6			
200	5	***2.6***	***6.0***	12.2	17.0	34.0	61.6	86.6	98.4	100.0
	10	***3.8***	***5.6***	13.8	23.4	49.8	84.4	98.2	100.0	100.0
	20	***3.2***	***6.0***	19.6	28.8	71.6	95.0	99.4	100.0	
	35	***3.4***	***7.8***	25.4	44.6	72.2	94.2	100.0	100.0	
	50	***4.8***	***13.0***	24.2	51.4	78.0	96.2	99.6	100.0	
350	5		***1.4***	9.8	10.8	21.2	44.2	73.0	91.8	99.8
	10		***3.4***	14.4	15.6	36.4	67.8	89.0	99.0	100.0
	20		***5.0***	14.2	23.2	53.4	82.0	95.0	100.0	100.0
	35		***2.0***	***19.2***	33.2	66.6	79.6	96.4	99.8	100.0
	50	***3.6***	***4.4***	***19.6***	41.2	64.4	82.6	97.2	100.0	100.0

Appendix 4: Correct Mixture Hypothesis Conclusion % (continued)

Note: Entries in emboldened italics correspond to correct treatment effect mean hypothesis conclusions below 100%.

$\sigma^2_{rat} = 2.00$

n_T	ϕ	3	4	7	10	15	20	25	30	40
100	5	*15.0*	*17.6*	25.8	45.0	72.2	94.0	99.2		
	10	*19.8*	*24.4*	43.4	62.4	87.6	99.6	99.8		
	20	*27.4*	*34.6*	55.2	75.4	96.4	99.8	100.0		
	35	*26.8*	*35.4*	*51.6*	71.6	96.2	100.0	100.0		
	50	*24.6*	*31.0*	*45.0*	68.2	96.6	100.0	100.0		
200	5	*17.0*	*15.6*	27.4	41.8	71.8	88.2	97.0	99.8	100.0
	10	*23.6*	*28.4*	38.2	63.8	89.0	98.8	99.6	100.0	100.0
	20	*32.8*	*38.8*	58.0	81.2	96.2	99.4	100.0	100.0	100.0
	35	*37.6*	*44.4*	68.8	82.6	95.8	99.8	100.0	100.0	100.0
	50	*33.4*	*39.2*	*58.8*	71.6	91.8	99.6	100.0	100.0	100.0
350	5		*20.8*	28.4	41.8	68.2	86.0	97.4	99.6	100.0
	10		*35.0*	50.4	63.6	87.4	97.6	99.6	100.0	100.0
	20		*45.4*	70.0	85.0	96.4	99.4	100.0	100.0	100.0
	35		*49.4*	74.4	89.2	97.8	99.6	100.0	100.0	100.0
	50		*48.0*	*66.6*	81.4	92.4	99.2	100.0	100.0	100.0

$\sigma^2_{rat} = 3.00$

n_T	ϕ	3	4	7	10	15	20	25	30	40
100	10	*39.6*	*43.6*	63.2	81.8	97.0	100.0	100.0		
	20	*56.8*	*62.6*	78.8	93.6	99.0	100.0	100.0	100.0	
	35	*56.4*	*63.8*	*85.8*	93.0	99.0	100.0	100.0	100.0	
200	10	*53.0*	*54.6*	71.8	82.6	98.2	99.6	100.0	100.0	100.0
	20	*65.4*	*72.8*	87.8	96.4	99.6	100.0	100.0	100.0	100.0
	35	*72.4*	*75.4*	91.2	97.0	99.6	100.0	100.0	100.0	100.0

Appendix 5: Correct Variance Hypothesis Conclusion %

Note: Entries in emboldened italics correspond to correct treatment effect mean hypothesis conclusions below 100%.

$\sigma^2_{rat} = 0.33$

						z*				
n_T	ϕ	3	4	7	10	15	20	25	30	40
100	0	*100.0*	99.6	88.8	94.4	100.0	100.0	100.0		
	10	*100.0*	*99.4*	87.0	92.8	98.8	99.8	100.0		
	20	*100.0*	*98.0*	80.2	92.6	98.8	99.6			
	35	*98.4*	*91.0*	73.8	84.0	95.6	99.0			
200	0	*100.0*	100.0	99.6	96.2	99.8	100.0	100.0	100.0	
	10	*100.0*	*98.8*	98.2	97.0	100.0	100.0	100.0	100.0	
	20	*100.0*	*100.0*	98.8	94.2	100.0	100.0	100.0	100.0	
	35	*100.0*	*100.0*	91.4	88.4	97.4	100.0	100.0	100.0	

$\sigma^2_{rat} = 0.50$

						z*				
n_T	ϕ	3	4	7	10	15	20	25	30	40
100	0	*97.8*	*89.0*	69.2	82.0	93.0	97.8	97.2		
	5	*97.0*	*87.2*	69.6	81.0	92.0	92.2	95.2		
	10	*93.8*	*86.6*	66.8	79.4	90.4	92.2	94.0		
	20	*92.6*	*80.4*	60.6	74.2	83.6	89.4	92.8		
	35	*86.0*	*69.6*	60.2	64.8	75.0	81.8			
	50	*73.8*	*57.0*	46.2	54.0	67.8	79.6			
200	0	*100.0*	100.0	87.4	81.4	97.2	99.2	99.8	100.0	
	5	*100.0*	*100.0*	84.6	79.6	94.0	99.2	99.6	100.0	
	10	*100.0*	*100.0*	82.2	81.6	94.6	98.4	99.2	99.4	
	20	*100.0*	*99.6*	77.6	77.6	90.4	96.8	99.0	99.6	
	35	*98.8*	*97.2*	64.8	65.4	85.2	94.6	96.0	98.0	
	50	*95.4*	*89.4*	51.4	58.0	76.0	87.8	93.8		
350	0		*100.0*	99.6	87.4	92.2	99.4	99.8	100.0	100.0
	5		*100.0*	100.0	90.2	93.6	99.2	100.0	100.0	100.0
	10		*100.0*	99.6	84.6	95.4	98.2	100.0	100.0	100.0
	20		*100.0*	98.6	81.0	90.6	98.0	100.0	100.0	100.0
	35		*100.0*	90.2	68.0	86.0	95.8	99.0	99.4	100.0
	50		*98.6*	72.4	58.0	81.6	90.6	96.4	99.0	

Appendix 5: Correct Variance Hypothesis Conclusion % (continued)

Note: Entries in emboldened italics correspond to correct treatment effect mean hypothesis conclusions below 100%.

$\sigma^2_{rat} = 1.0$

n_T	ϕ	3	4	7	10	15	20	25	30	40
50	0	***82.6***	***85.2***	83.4	82.4	85.4	85.0			
	5	***82.4***	***84.0***	84.0	79.4	79.8	84.4			
	10	***82.2***	***84.6***	78.6	80.2	84.8	83.2			
	20	***79.4***	***84.4***	73.8	72.2	79.8				
	35	***80.2***	***80.4***	71.2	64.6	82.0				
	50	***79.4***	***76.8***	***62.0***	70.0	83.6				
100	0	***85.6***	***84.8***	84.8	83.4	79.8	83.0	83.8		
	5	***83.8***	***80.8***	88.4	82.6	81.6	81.0	84.8		
	10	***80.8***	***85.0***	83.0	80.0	77.0	84.4	83.4		
	20	***80.0***	***83.0***	80.6	73.8	76.2	83.6	79.6		
	35	***83.0***	***85.6***	71.2	66.4	73.8	86.4	81.0		
	50	***80.8***	***82.2***	***70.8***	65.2	74.4	83.0			
200	0	***82.2***	***85.4***	85.0	84.6	83.6	84.2	85.0	83.6	88.0
	5	***82.8***	***84.6***	88.4	84.0	81.0	77.2	80.0	85.6	80.2
	10	***79.8***	***80.0***	85.4	79.0	74.4	81.0	84.6	87.2	81.2
	20	***80.0***	***82.6***	83.0	77.2	71.6	84.0	83.6	83.0	
	35	***82.2***	***82.0***	79.4	69.8	65.4	83.0	84.4	82.6	
	50	***79.6***	***81.2***	73.2	58.2	64.2	82.4	83.8	81.8	
350	0		***83.0***	85.2	83.6	86.2	83.8	86.4	86.0	82.4
	5		***83.4***	86.2	83.2	81.2	79.8	79.0	81.0	82.2
	10		***80.0***	85.8	84.0	79.0	73.4	81.8	82.4	80.0
	20		***82.8***	83.0	79.0	69.8	74.0	82.0	83.4	83.8
	35		***83.0***	***78.6***	71.0	64.0	71.2	83.2	84.8	85.4
	50	***82.4***	***81.6***	***78.2***	66.6	56.8	72.0	84.8	87.4	84.4

Appendix 5: Correct Variance Hypothesis Conclusion % (continued)

Note: Entries in emboldened italics correspond to correct treatment effect mean hypothesis conclusions below 100%.

$\sigma^2_{rat} = 2.00$

n_T	ϕ	3	4	7	10	15	20	25	30	40
						z*				
100	0	*96.8*	*97.6*	97.2	97.2	98.8	98.8	98.0		
	5	*96.4*	*95.0*	96.2	94.2	95.0	95.8	96.2		
	10	*93.8*	*92.0*	92.8	88.4	91.0	95.2	95.6		
	20	*89.4*	*86.4*	81.8	83.4	91.4	91.8	93.4		
	35	*72.8*	*67.6*	*67.2*	71.8	84.8	87.4	90.6		
	50	*55.8*	*53.2*	*52.8*	62.0	76.2	75.8	84.0		
200	0	*100.0*	*100.0*	100.0	100.0	100.0	100.0	100.0	100.0	
	5	*100.0*	*100.0*	99.8	99.8	99.6	99.8	99.8	99.8	100.0
	10	*100.0*	*99.6*	99.4	99.8	98.8	99.6	99.8	100.0	100.0
	20	*98.6*	*98.4*	98.6	98.6	99.0	99.0	99.6	99.6	99.8
	35	*91.2*	*90.6*	91.4	91.0	93.8	99.2	98.4	99.2	99.6
	50	*79.0*	*78.0*	*73.8*	75.8	86.6	93.0	93.4	94.6	96.8
350	0		*100.0*	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	5		*100.0*	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	10		*100.0*	100.0	100.0	99.8	100.0	100.0	100.0	100.0
	20		*100.0*	99.8	100.0	99.8	99.8	100.0	100.0	100.0
	35		*98.4*	98.0	98.6	98.6	99.6	100.0	100.0	100.0
	50		*94.0*	*90.0*	91.2	93.2	97.6	99.8	98.6	99.4

$\sigma^2_{rat} = 3.00$

n_T	ϕ	3	4	7	10	15	20	25	30	40
						z*				
100	0	*100.0*	100.0	100.0	100.0	100.0	100.0	100.0		
	10	*100.0*	*100.0*	99.8	100.0	100.0	100.0	100.0		
	20	*99.6*	*99.2*	99.8	99.6	99.4	99.8	100.0	99.8	
	35	*95.2*	*95.4*	*94.6*	96.0	98.6	99.4	99.6	99.8	
200	0	*100.0*	*100.0*	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	10	*100.0*	*100*	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	20	*100.0*	*100*	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	35	*100.0*	*99.8*	99.8	100.0	99.8	100.0	100.0	100.0	100.0

References

Akaike, H. (1973). Information Theory and an extension of the maximum likelihood principle. In B.N. Petrov & B.F. Csaki (Eds), *Second International Symposium on Information Theory*, 267-281. Academiai Kiado: Budapest.

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics, 16*, 3-14.

Akaike, H. (1977). On entropy maximization principle. In P.R. Krishnaiah (Ed.), *Proceedings of the Symposium on Applications and Statistics, 27-47*. Amsterdam: North Holland.

Anderson, T.W. & Amemiya, Y. (1988). The asymptotic normal distribution of estimators in factor analysis under general conditions. *Annals of Statistics, 16*, 759-771.

Angrist, J.D., Imbens, G.W., & Rubin, D.B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*, 444-455.

Aptech Systems. (2005). *Gauss Mathematical and Statistical System: Version 8*. Maple Valley, WA: Aptech Systems Inc.

Bentler, P. M. (2006). *EQS Structural Equations Program Manual.* Encino, CA: Multivariate Software, Inc.

Bollen, K.A. 1989. *Structural Equations and Latent Variables*. John Wiley & Sons Inc., New York.

Boos, D., & Brownie, C. (1991). Mixture models for continuous data in dose-response studies when some animals are unaffected by treatment. *Biometrics, 47*, 1489-1504.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika, 52*, 345-370.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences. 2nd Edition*. Hillsdale, NJ: Erlbaum.

Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1993). Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. *Psychological Bulletin, 114,* 174-184.

Conover, W.J. (1999). *Practical Nonparametric Statistics, 3rd ed*. New York, NY: John Wiley & Sons, Inc.

Coombs, W.T., Algina, J., & Oltman, D. (1996). Univariate and multivariate omnibus hypothesis tests to control type I error rates when population variances are not necessarily equal. *Review of Educational Research, 66*, 137-79.

Curran, P.J., West, S.G., & Finch, J.F. (1996). The robustness of test statistics to non-normality and specification error in confirmatory factor analysis. *Psychological Methods, 1*, 16-29.

D'Agostino, R. and Stephens, M., eds (1986). *Goodness-of-Fit Techniques*. New York: Marcel Dekker, Inc.

Dayton, C.M. (2003a). Information criteria for pairwise comparisons. *Psychological Methods, 8*, 61-71.

Dayton, C.M. (1998). Information criteria for the paired-comparison problem. *The American Statistician, 52*, 144-151.

Dayton, C.M. (2003b). Model comparisons using information measures. *Journal of Modern Applied Statistical Methods, 2*, 281-292.

Dempster A., Laird N., & Rubin D. (1977). Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society Series B, 39*, 1-38.

Devore, J.L. (2000) *Probability and Statistics for Engineering and the Sciences, 5th Edition*. Pacific Grove, CA: Duxbury.

Doornik, J. A., & Hansen, H. (1994). *An omnibus test for univariate and multivariate normality* (Working paper). NuOEeld College, Oxford.

Everitt, B.S., & Hand, D.J. (1981). *Finite mixture distributions*. London: Chapman & Hall.

Gans, D.J. (1984). The search for significance: different tests on the same data. *Journal of Statistical Computation and Simulation, 19*, 1-21.

Gray, G. (1994). Bias in misspecified mixtures. *Biometrics, 50*, 457-470.

Grun B., & Leisch F. (2004). Bootstrapping Finite Mixture Models. In J Antoch (ed.), *Compustat 2004 – Proceeding in Computational Statistics*, pp 1115-1122. Physika Verlag, Heidelberg, Germany.

References (continued)

Hancock, G. R. (2003). Fortune cookies, measurement error, and experimental design. *Journal of Modern Applied Statistical Methods*, *2*, 293-305.

Hancock, G. R. (2004). Experimental, quasi-experimental, and nonexperimental design and analysis with latent variables. In D. Kaplan (Ed.), *The SAGE Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks, CA: SAGE Publications.

Hannan, E.J. & Quinn, B.G. (1979). The determination of an order of an autoregression. *Journal of the Royal Statistical Society Series B, 41*, 190-195.

Hathaway, R.J. (1985). A constrained formulation of the maximum likelihood estimation for normal mixture distributions. *The Annals of Statistics, 13*, 795-800.

Haughton, D. (1997). Packages for estimating finite mixtures: A review. *The American Statistician, 51*, 194-205.

Hollis, S. & Campbell, F. (1999). What is meant by intent to treat analysis? Survey of published randomized controlled trials. *British Medical Journal, 319*, 670-674.

Huang, C.C. & Dayton, C.M. (1995). Detecting patterns of bivariate mean vectors using model-selection criteria. *British Journal of Mathematical & Statistical Psychology, 48*, 129-147.

Kaplan, D. (2000). *Structural equation modeling*. Thousand Oaks, CA: Sage.

Keppel, G. & Wickens, T.D. (2004). *Design and Analysis: A Researcher's Handbook. 4th Edition*. Englewood Cliffs, NJ: Prentice Hall.

Khamis, H.J. (2000). The two-stage delta-corrected Kolmogorov-Smirnov test, *Journal of Applied Statistics, Vol. 27, 4,* 439-450.

Kirk, R.E. (1995). *Experimental Design. 3rd Edition*. Pacific Grove, CA: Brooks/Cole.

Kolenikov, S. (2001). Decomposition of normal mixture by maximum likelihood:denormix package. *Stata NASUG*, 2001.

Kullback, S. & Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*, 79-86.

References (continued)

Kullback, S. (1959). *Information Theory and Statistics*. New York: John Wiley & Sons.

Lachin, J.M. (2000). Statistical considerations in the intent to treat principle. *Controlled Clinical Trials, 21*, 167-189.

Larimore, W.E. & Mehra, R.K. (1985). The problems of overfitting data. *Byte*, 167-180.

Leeb, H. & Pötcher, B. (2005). Model selection and inference: fact and fiction. *Econometric Theory, 21*, 21-59.

Leeb, H. & Pötcher, B. (2006). Can one estimate the unconditional distribution of post-model-selection estimators? *Annals of Statistics*, *34*, 2554-2591.

Leemis, L.T. & McQueston, J.T (2008). Univariate Distribution Relationships. *The American Statistician, 62*, 45-53.

Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software, 11*, 1-18.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of Econometrics, 16*, 3-14.

Luo, X., Boos, D.D., & Tamura, R.N. (2004). Score tests for dose effect in the presence of non-responders. *Journal of the Royal Statistical Society, Series B (Methodological), 44, 2*, 226-233.

Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57, 3*, 519-530.

Mardia, K.V. (1980). Tests of univariate and multivariate normality. In *Handbook in Statistics*, Ed. P. R. Krishnaiah, 279-320. Amsterdam: North-Holland.

McLachlan, G. J. & Peel, D. (2000). MIXFIT: An algorithm for the automatic fitting and testing of normal mixture models. *Proceedings of the 14th International Conference on Pattern Recognition*, Vol. I, Los Alamitos, CA: IEEE Computer Society, 553-557.

McLachlan, G.J. & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Moser, B.K. & Stevens, G.R. (1992). Homogeneity of Variance in the Two-Sample Means Test. *The American Statistician, 46*, 19-21.

References (continued)

Mudholkar, G.S., McDermott, M., & Srivastava, D.K (1992). A test of p-variate normality. *Biometrika*, *79, 4*, 850-854.

Muthén, B., & Muthén, L. (2001). *Mplus User's Guide*. Los Angeles, CA: Muthén and Muthén.

Olsson, U., Troye, S.V., & Howell R.D. (1999). Theoretical fit and empirical fit: The performance of maximum likelihood versus generalized least squares estimation in structural equation models. *Multivariate Behavorial Research, 34*, 31-58.

Park, H.M. (2008). Comparing group means: T-tests and one-way ANOVA Using Stata, SAS, and SPSS. The Trustees of Indiana University. http://www.indiana.edu/~statmath/stat/all/ttest/ttest.pdf

Pavlic, M., Brand, R. J., & Cumming, S. R. (2001). Estimating probability of nonresponse to treatment using mixture distributions. *Statistics in Medicine, 20*, 1739-1753.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing; Version 2.6.1.* R Foundation for Statistical Computing, Vienna, Austria.

Radolfi A., & Idier J. (1999). Penalized maximum likelihood estimation for univariate normal mixture distributions. In *Actes du 17 Colloque GRETSI*, 259-262.

Satorra, A., & Bentler, P.M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C.C. Clogg (Eds), *Latent variables analysis: Applications for developmental research*. Thousand Oaks, CA: Sage.

Satorra, A. (2003). Power of chi-square goodness-of-fit test in structural equation models: the case of non-normal data. In Okada, A., Shigemasu, K., Kano, Y. & Meulman, J.J. (Eds.), *New Developments of Psychometrics.* Springer Verlag: Tokyo.

Satterwaite, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2*, 110-114.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Sheskin, D. (2007). *Handbook of parametric and nonparametric statistical procedures, 4th ed.* Boca Raton, FL: Chapman & Hall.

Shibata, R. (1983). A theoretical view of the use of AIC. In O.D. Anderson (Ed.), *Time Series Analysis: theory and Practice, 4*, 237-244. Amsterdam: North Holland.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27,* 229-239.

StataCorp. (2007). *Stata Statistical Software: Release 10.* College Station, TX: StataCorp LP.

Titterington D., Smith A.,& Makov U. (1985). *Statistical Analysis of Finite Mixture Distributions.* Chichester: Wiley.

Wedel M. & DeSarbo W.S. (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification, 12*, 21-55.

Welch, B.L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 29*, 350-362.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica, 50,*1-25.

Zimmerman, D.W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology, 57*, 173-81.