The Comparison of Treatments with Ordinal Responses

LU, Tongyu

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy
in
Statistics

The Chinese University of Hong Kong
June 2011

Thesis/Assessment Committee

Professor CHEUNG Siu Hung (Chair)

Professor POON Wai-Yin (Thesis Supervisor)

Professor WU Ka-Ho (Committee Member)

Professor HU Feifang (External Examiner)

Abstract

The comparison of treatments to detect possible treatment effects is a very important topic in statistical research. It has been drawing significant interests from both academicians and practitioners. Important research work on treatment comparisons dates back several decades. For treatment comparisons, the following three cases are very common: the comparison of two independent treatments; the comparison of treatments with repeated measurements; and the multiple comparison of several treatments. For different cases, the involved research issues are usually different. In many fields of study, the level of measurement for responses of the treatments is ordinal. Many examples can be found in areas such as biostatistics, psychology, sociology, and market research, where the ordered categorical variables play an important role.

In this thesis, we focus on the the comparison of treatments with ordered categorical responses. The three cases of treatment comparisons will all be studied. The main objective of this thesis is to develop more effective comparison methods for treatments with ordinal responses and to address some important issues involved in different comparison problems. Our major statistical approach is to consider ordinal responses as manifestations of some underlying continuous random variables.

This thesis consists of three main parts. In the first part, we consider the modeling of treatments with longitudinal ordinal responses by a latent growth curve. On the basis of such a latent growth curve, we achieve a comprehensive flexible model with straightforward interpretations and a variety of applications including treatment comparison, the analysis of covariates, and equivalence test of treatments. In the second part, we consider the comparison of several treatments with a control for ordinal responses. By considering the ordinal responses as manifestations of some underlying normal random variables, a latent normal distribution model is utilized and the corresponding parameter estimation method is proposed. Further, we also derive testing procedures that compare several treatments with a control

under an analytical framework. Both single-step and stepwise procedures are introduced, and these procedures are compared in terms of average power based on a simulation study. In the last part of this thesis, we establish a unified framework for treatment comparisons with ordinal responses, which allows various treatment comparison methods be comprehended using a unified perspective. The latent variable model is also utilized, but the underlying random variables are allowed to have any member of the location-scale distribution family. This latent variable model under such a specification of underlying distributions subsumes many existing models in the literature. A two-step procedure to identify the model and produce the parameter estimates is proposed. Based on this procedure, many important statistical inferences can be conveniently conducted. Furthermore, the sample size determination method based on the latent variable method is also proposed. The proposed latent variable method is compared with the existing methods in terms of power and sample size.

摘要

通過對處理的比較來檢測可能存在的處理效應是統計研究中一個非常重要的課題。這一問題已經引起了理論研究人員和實際應用人員的廣泛興趣。對於處理比較的重要研究可以追溯到幾十年之前。它包括以下三類重要的類型：對兩個獨立處理的比較；對重複觀測處理的比較；以及對多個處理的多重比較。對於不同的類型，它們所包含的研究問題通常是不一樣的。在很多的研究領域，對處理的響應變量通常採用有序分類的度量尺度。在許多領域，例如生物醫學、心理學、社會學、以及市場研究中都能發現大量的實例，而且有序分類響應變量在這些領域中都扮演著重要的角色。

在本論文中，我們將重點研究具有有序分類響應變量的處理的比較問題。對於處理比較的三種重要類型都有研究。 本論文的目標是對具有有序分類響應變量的處理提出更爲有效的比較方法， 同時註意解決在不同類型的研究中所包含的問題。我們的一個主要的統計方法是把有序分類響應變量看作是某些潛在連續變量的一種表現。

本論文主要由三部分構成：在第一部分， 我們主要研究用潛成長曲線來對重複觀測的具有有序分類響應變量的處理的建模問題。我們構建的潛成長曲線模型具有非常直觀的解釋。基於這一模型，我們可以進行多種的分析和應用，這包括處理比較、協變量分析、以及處理等價性檢驗等。在論文的第二部分，我們主要考慮多個處理對一個控制處理的多重比較問題。通過把有序分類變量看作是潛在正態變量的一種表現，我們構造了正態潛變量模型並提出了相應的參數估計方法。基於這一正態潛變量模型，我們給出了用來比較多個處理對一個控制處理的若干檢驗過程，包括單步檢驗過程和逐步檢驗過程。並通過隨機模擬，對這些檢驗過程在檢驗的平均功效方面進行了比較分析。在論文的第三部分，我們對具有有序分類響應變量的處理的比較問題提出了一個統一的分析框架。在這一框架下，可以對各種不同的處理比較方法從一個共同的角度來認識。我們仍然採用潛變量模型的分析方法，但是允許這些潛變量具有位置-尺度分佈族中的任意分佈。具有這種分佈假定的潛變量模型包括了許多這一背景下的重要模型。爲了模型的識別和參數估計，我們提出了一個兩步估計過程。基於這一估計過程，很多重要的統計推斷可以很方便的進行。 此外，我們還給出了基於潛變量模型的樣本量的確定方法。對於我們提出的潛變量方法，我們從檢驗的功效和樣本量的大小兩個方面與已有的方法進行了比較。

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Prof. Wai-Yin Poon, for her support, guidance and encouragement throughout my PhD studies. In the process of preparing and writing research papers and thesis, Prof. Poon gave me a lot of valuable suggestions and insights that greatly benefit my research. I am also extremely grateful to Prof. Siu Hung Cheung who led me to the interesting research area of multiple testing and is always willing to help and to give his best suggestions.

Special thanks go to the members of my thesis committee, Prof. Feifang Hu, Prof. Ka-Ho Wu, and Prof. Siu Hung Cheung, for their time and efforts. I would also like to thank Prof. Anthony Hayter for his valuable suggestions on an earlier version of a manuscript based on material in chapter 3 in this thesis.

I am also grateful to all kinds of help from the professors and staff in the Department of Statistics. I would also like to express thanks to all current and former graduate students for their help and support. The friendship from all of you will be cherished forever.

Finally, I am deeply indebted to my wife, Jihong Lai, and my parents for their love and care through these years. This thesis is dedicated to them.

Contents

1 Introduction **1**

1.1 Background and Literature Review 1

1.2 The Objective and Outline 4

2 Latent Growth Curve Modeling of Longitudinal Ordinal Responses **7**

2.1 Introduction 7

2.2 Latent Variable Growth Curve Modeling of Ordinal Data 10

2.3 Comparing Two Treatments 15

2.4 Other Applications of the LCM 21

2.4.1 Analyzing the Effects of Covariates 21

2.4.2 Inference for Equivalence Treatments 25

2.5 Conclusion 27

3 Multiple Testing of Several Treatments with a Control **29**

3.1 Introduction 29

3.2 The Latent Variable Model 31

3.3 The Proportional Odds Model 33

3.4 Parameter Estimation of the Latent Variable Model 37

3.5 Multiple Testing of Several Treatments with Control 40

3.5.1 Test Statistics 40

3.5.2 Multiple Testing Procedures 41

3.5.3 Power Comparison: a Simulation Study 44

3.6 Examples 48
3.6.1 Example 1 48
3.6.2 Example 2 50
3.7 Conclusion 52

4 A Unified Framework for Treatment Comparisons **54**
4.1 Introduction 54
4.2 A Unified Consideration of the Treatment Effect Measures 57
4.3 A Framework for the Analysis of the Latent Variable Model 61
4.3.1 Model Estimation 61
4.3.2 Statistical Inference 66
4.4 The Power and Sample Size Determination in the Comparison of Two Independent Treatments 70
4.4.1 The Existing Methods 70
4.4.2 The Latent Variable Method 73
4.4.3 The Comparison of Different Methods 76
4.5 A Real Data Example 80
4.6 Further Study on Sample Size Determination 83
4.7 Conclusion 87

5 Future Research **88**

Bibliography **94**

A The Mx Input Script for LCM **102**

B The Proof of Theorem 3.2 and Theorem 3.3 **108**

C The Proof of Lemma 4.1 **113**

D The Score Function and Fisher Information Matrix for the Likelihood Function (4.14) and (4.22) **115**

List of Figures

2.1 Growth curves 12

2.2 The latent growth curve for the active drug and placebo treatments . 18

3.1 Average power of H, DTSD, DSS, and B. 45

3.2 Increase in average power of H, DTSD, and DSS as compared to B. . 46

3.3 Average power of H, DTSD, DSS, and B for different true/false configurations. 48

3.4 All-pairs power of H, DTSD, DSS, and B for different true/false configurations. 49

List of Tables

2.1 Time to fall asleep (in minutes), by treatment and occasion. 8
2.2 Estimates of the model parameters for the active drug group 14
2.3 Maximum likelihood estimates of the model parameters 17
2.4 Summary of tests 21
2.5 Parameters estimates in the LCM with a dummy variable covariate . 24

3.1 Ordered categorical data of a clinical trial. 32
3.2 Estimated rejection rate (E) of the null hypothesis 37
3.3 Incidence and intensity of pain on injection of propofol 50
3.4 Incidence and intensity of pain on injection of propofol 51

4.1 Ordinal data of two treatments organized in contingency table 70
4.2 The power of different testing methods. 77
4.3 The sample size determination of different methods. 79
4.4 Retinopathy category by smoking status for 613 diabetic patients. . . 80
4.5 The determination of sample size for the real example 82
4.6 The sample size determination of different methods for given ordinal responses 86

Chapter 1

Introduction

1.1 Background and Literature Review

Treatment comparison is a very important topic in statistical research, since in many practical studies, the comparison of treatments to detect possible treatment effects is usually one of the most important issues. In many fields of study where precise measurement is not possible, the responses of the treatments are usually measured in an ordinal scale. For example, in a medical study, the recovery status of the patients may be evaluated as "good," "satisfactory," "bad," and "very bad". In many other area, such as psychology, sociology, and market research, the ordered categorical variables also play important roles.

Different methods of analyzing ordinal categorical data are available in the literature. Classical methods include those used to compare the ordinal responses obtained from two independent samples, such as the log-linear-type row effect model and the Mann-Whitney statistic (Agresti, 1984); those that investigate the association between two ordinal categorical variables that are cross-classified into a contingency table, such as the uniform association model and the mean response model (Agresti, 1983); and those that analyze the effects of covariates, such as the linear logistic model for dichotomous data suggested by Cox (1970) and the proportional odds or proportional hazards models discussed by McCullagh (1980).

For treatment comparison, the following three cases are common:

TC1: the comparison of two independent treatments;

TC2: the comparison of treatments with repeated measurements at different time points; and

TC3: the multiple comparison of several treatments.

For different cases, the emphases of research are usually different. In this thesis, we are interested in the comparison of treatments with ordered categorical responses. A brief literature review of the three cases of treatment comparison for ordinal response is given as follows.

TC1: the comparison of two independent treatments.

The comparison of two independent treatments (or a treatment and a control) should be the most fundamental task in treatment comparison study. When the responses of the treatments are continuous, the classical t test can be used to examine the mean difference between two treatments based on the normality assumption. The Wilcoxon-Mann-Whitney (WMW) test (Wilcoxon 1945, Mann and Whitney 1947) may be the most popular nonparametric method that is used to investigate the effect between two treatments with ordered categorical responses. As a distribution-free method, the WMW test is also widely used to compare treatments with continuous responses. A comprehensive study of the WMW test is given by Lehmann (1975).

For ordered categorical response, it is in many cases reasonable to consider the ordinal response as the manifestation of an underlying continuous variable (see e.g. McCullagh, 1980; Anderson, 1984). The proportional odds model proposed by McCullagh (1980) has been widely adopted in the literature. To compare treatments with ordered categorical responses, Whitehead (1993) proposed to use the log-odds ratio as a measure of mean treatment effect under the proportional odds assumption, and used the WMW test to investigate the treatment effect. By considering the ordinal response as the manifestation of underlying normally distributed random variable, Poon (2004) proposed a method to examine the possible treatment effect,

which can be conveniently attributed to either location effect or dispersion effect based on the latent variable model.

TC2: the comparison of treatments with repeated measurements.

When the treatments are measured in a longitudinal manner. Besides the treatment effects, the time effects or the growth trends of the treatments effects becomes another interest of study. For the study of longitudinal data, some important work has been achieved, which usually focuses on measuring the association between the responses and the covariates (see e.g. Liang and Zeger, 1986; Zeger and Liang, 1986; Prentice, 1988).

The commonly used methods for modeling longitudinal ordinal categorical responses can be broadly categorized into three classes. The first class comprises marginal models that focus on the modeling of marginal probabilities. The most popular are the cumulative-type models (Agresti, 1999, 2002; McCullagh, 1980) such as the cumulative probit and cumulative logit models. The second class includes the Markov chain transitional models, which focus on the modeling of transition probabilities (Kalbfleisch and Lawless, 1985; Chan and Munoz-Hernandez, 2003). The third class prefers to use the latent variable model for such longitudinal ordinal responses (see e.g. Qu, Piedmonte, and Medendorp, 1995; Todem, Kim, and Lesaffre, 2007), where the time dependent effects are measured by the correlation structure of the underlying continuous variables.

TC3: the multiple comparison of several treatments.

Multiple comparison of treatments has a long research history. Two main types of multiple comparisons are usually studied in the literature, the comparison of several treatments with a control (see e.g. Dunnett, 1955; Dunnett and Tamhane, 1991, 1992), and the pairwise comparison (see e.g. Tukey, 1953; Hayter, 1984, 1989). A wealth of multiple comparison procedures are also available, see e.g. the excellent review books by Hochberg and Tamhane (1987), Hsu (1996).

In multiple comparison, a fundamental task is the control of family-wise error (FWE), which is defined as the probability of rejecting any true null hypothesis.

Both single-step and stepwise procedures have been proposed in order to improve the power of tests, see e.g. the step-down procedure proposed by Holm (1979), and the step-up procedure proposed by Hochberg (1988). There are several possible definitions of power in the literature on multiple comparison. Horn and Dunnett (2004) discussed several different definitions, such as, all-pairs power, any-pair power, per-pair power, and average power. The existing multiple comparison procedures usually focus on the comparison of treatments with continuous responses. Little work has been done on the multiple comparison of treatments with ordinal responses.

In treatment comparison, the determination of sample size to achieve a specified power level is also an important issue, especially in the planning stage of an experiment. In the literature of comparing two treatments with ordered categorical responses, several sample size determination methods have been proposed. For example, Whitehead (1993) provided a sample size formula that is derived based on the WMW test with the alternative specified as proportional odds. Zhao, Rahardja, and Qu (2008) gave the sample size calculation for the WMW test with the alternative specified as the probability of one treatment being superior to the other. However, for the case of multiple comparison of treatments with ordinal responses, little work on sample size determination has been done.

1.2 The Objective and Outline

Our study focuses on the comparison problem of treatments that have ordered categorical responses. The aforementioned three cases of treatment comparisons will all be studied in this thesis. The main objective of this thesis is to develop effective comparison method and to address some important issues involved in different comparison problems. The main idea of our methods is to consider the ordered categorical responses as manifestations of some underlying continuous random variables. On the basis of such latent variable model, the treatments with ordinal responses can be characterized by the corresponding underlying distributions, and many statistical

inferences can be conducted conveniently.

A brief summary of the subsequent chapters is outlined as follows.

Chapter 2: In this chapter, we consider the use of the latent growth curve model to analyze longitudinal ordinal categorical data that involve measurements at different time points. By operating on the assumption that the ordinal response variables at different time points are related to normally distributed underlying continuous variables, and by further modeling these underlying continuous variables for different time points with the latent growth curve model, we achieve a comprehensive and flexible model with straightforward interpretations and a variety of applications. We discuss the applications of the model in treatment comparisons and in the analysis of the covariate effects. Moreover, one prominent advantage of the model lies in its ability to address possible difference in the initial conditions of the subjects who take part in different treatments. Making use of this property, we also develop a new method to test the equivalence of two treatments that involve ordinal responses obtained at two different time points. A real data set is used to illustrate the applicability and practicality of the proposed approach. The results described in this chapter have been summarized in the paper by Lu, Poon, and Tsang (2011).

Chapter 3: In this chapter, we consider the multiple comparison of several treatments with a control for ordered categorical responses. A motivation for our research is the study on the behavior of the WMW test. Our study finds that the level of the WMW test can not be preserved when the treatments differ in dispersion. We propose an alternative approach that can address this issue. By considering the ordinal responses of different treatments as manifestations of some underlying normal random variables, a latent normal distribution model is utilized and the corresponding parameter estimation method is proposed. Under the latent variable model framework, we derive testing procedures that compare several treatments with a control. Both single-step and stepwise procedures are introduced, and these procedures are compared in terms of average power based on simulation study. Multiple testing procedures for practical application are also suggested. Data from clinical trials are

used to illustrate the proposed procedures. The results described in this chapter have been summarized in the paper by Lu, Poon, and Cheung (2011).

Chapter 4: The theoretical method proposed in Chapter 3 is generalized to a more general case. The latent variable model is also utilized to analyze ordinal data in this chapter, but the underlying variables are allowed to have any distributions belonging to the location-scale family. On the basis of the proposed two-step estimation procedure, the locations and scales characterizing different treatments can be freely estimated. Consequently, many statistical inferences can be conveniently conducted based on the proposed methods. This analysis framework for ordinal data includes the mostly adopted models, such as the ordinal logistic model and the ordinal probit model, in the literature as special cases. Based on such an analysis framework the existing treatment effect measures for ordinal responses can be interpreted in a unified manner. Two important latent variable methods for treatment comparison, the *LNorm* method and the *LLogis* method, are detailed illustrated for the comparison of two treatments with ordinal responses. The corresponding sample size determination methods are also proposed, which can accommodate the difference in the scales of different treatments. The proposed methods are compared with the existing methods in terms of power and sample size by both numerical study and real example.

Chapter 5: This chapter concludes the thesis by listing some possible areas for future research. The improvement of the two-step estimation procedure proposed in Chapter 4 is also discussed as a remark.

Chapter 2

Latent Growth Curve Modeling of Longitudinal Ordinal Responses

2.1 Introduction

In this chapter, we consider the modeling and analysis of longitudinal ordinal responses that involve measurements at two different time points. We are specifically interested in the analysis of the type of data presented in Table 2.1, which is taken from Agresti (1989). In a double-blind clinical trial, an active hypnotic drug and a placebo were randomly administered to two independent samples of patients with insomnia. Each individual was asked at the start and end of a two-week treatment period the question: "How quickly did you fall asleep after going to bed?" The responses were classified into one of four categories: "< 20," "20-30," "30-60," and "> 60" (in minutes). For this data set, several research questions are of interests, including

(Q1) whether the initial conditions of the subjects receiving the two different treatments are the same;

(Q2) whether there is a significant efficacy difference between the active drug and the placebo;

(Q3) the direction and extent of the treatment effect of the active drug; and

(Q4) whether the two treatments are equivalent.

Treatment	Initial occasion	Follow-up occasion			
		< 20	$20 - 30$	$30 - 60$	> 60
Active drug	< 20	7	4	1	0
	$20 - 30$	11	5	2	2
	$30 - 60$	13	23	3	1
	> 60	9	17	13	8
Placebo	< 20	7	4	2	1
	$20 - 30$	14	5	1	0
	$30 - 60$	6	9	18	2
	> 60	4	11	14	22

Table 2.1: Time to fall asleep (in minutes), by treatment and occasion.

In this chapter, we proposed the use of the latent growth curve model (LCM) to analyze the data. The proposed method can provide answers to all these questions in a comprehensive manner, while commonly used existing methods can only address some of these questions.

The commonly used methods for modeling longitudinal ordinal categorical responses can be broadly categorized into two classes (Chan and Munoz-Hernandez, 2003). The first class comprises marginal models that focus on the modeling of marginal probabilities. The most popular are the cumulative-type models (Agresti, 1999, 2002; McCullagh, 1980) such as the cumulative probit and cumulative logit models. The second class includes the Markov chain transitional models, which focus on the modeling of transition probabilities (Kalbfleisch and Lawless, 1985; Chan and Munoz-Hernandez, 2003). In this chapter, we propose the use of the latent growth curve model (LCM). The use of LCM in analyzing continuous variables has been widely discussed in the literature (Duncan *et al.*, 1999; Bollen and Curran, 2006), but its use in modeling and analyzing ordinal categorical data has received little attention.

Like for the probit model, we assume that the ordinal response variables at different time points are related to normally distributed underlying continuous variables.

By using the LCM to further model the underlying continuous variables for different time points, we achieve a comprehensive and flexible model. This model has easy and straightforward interpretations, can be applied to analyze various types of medical data, can effectively compare the effects of two treatments whether or not the initial conditions of the subjects who receive those treatments are the same, and can be implemented in a number of easily accessible software programs. The general model framework facilitates different model generalization directions, and the model also has a variety of applications. We discuss the use of the model in treatment comparisons and in covariate analysis, and we also examine its applications in equivalence tests.

Establishing the equivalence of a newly developed treatment to a standard treatment is of interest in many medical studies. Equivalence can be established by showing that the responses to the two treatments differ by an insignificant and clinically acceptable amount. This method is particularly useful when a newly developed treatment is less expensive, easier to administer, or has fewer side effects. Statistical methods for the inference of equivalence treatments are widely available in the literature (e.g., Dunnett and Gent,1977; Nam, 1997; Lui and Cumberland, 2001; Liu *et al.*, 2002; Lui and Zhou, 2004; Wang *et al.*, 2006; Tang and Poon, 2007). More specifically, Lui and Cumberland (2001) developed an equivalence test for ordinal data with matched-pairs that can be organized in a contingency table, and used the marginal proportions to assess the equivalence of two treatments. Tang and Poon (2007) considered two independent treatments with ordinal responses, and used a latent normal distribution approach to establish the equivalence of two treatments. However, these methods focus on comparing the responses of two treatments, and fail to address possible differences in the initial conditions of the test subjects. In other words, these methods are not optimal if they are used to analyze data with measures of initial conditions, such as the data given in Table 2.1. Based on the proposed LCM, an equivalence test for ordinal responses, which can effectively address possible differences in the initial conditions of subjects, is developed. The data set

in Table 2.1 is analyzed as an illustration.

The structure of this chapter is as follows. In Section 2.2, we introduce a latent variable growth curve for modeling ordinal responses with measures at two different time points, and discuss the maximum likelihood estimation method. In Section 2.3, we discuss how the model can be used in a flexible and effective manner to compare two treatments. In Section 2.4, we discuss further applications of the LCM. We also generalize the basic model to analyze the effects of covariates and develop a procedure for the statistical inference of equivalence tests. Section 2.5 concludes this chapter with a discussion.

2.2 Latent Variable Growth Curve Modeling of Ordinal Data

Let y_{it}^* be the observed ordinal variables for individual i at time t, $t = 1, 2, \cdots, T$. We operate on the assumption that it is related to an underlying continuous variable, y_{it}, via a latent variable model given by

$$y_{it}^* = k \quad \text{iff} \quad \tau_{k-1} < y_{it} \leq \tau_k, \tag{2.1}$$

where $k = 1, 2, \ldots, K$, $-\infty = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_{K-1} < \tau_K = +\infty$. With K response categories, there are $K - 1$ threshold parameters, $\tau_k, k = 1, 2, \ldots, K - 1$. Although dealing with different numbers of response categories at different time points is theoretically straightforward, it is sensible to assume that the total number of response categories are the same at different time points when the same variable is measured repeatedly. To examine the trend in the ordinal variable, we consider that the underlying continuous variables at different time points t can be described

by a latent curve model (LCM) that is given by the following set of equations.

$$\begin{aligned} \text{Trajectory equation:} \quad & y_{it} = \alpha_i + \lambda_t \beta_i + \epsilon_{it} \\ \text{Intercept equation:} \quad & \alpha_i = \mu_\alpha + \zeta_{\alpha i} \\ \text{Slope equation:} \quad & \beta_i = \mu_\beta + \zeta_{\beta i} \end{aligned} \tag{2.2}$$

This set of equations represent a very general model and it is necessary to impose constraints on the relevant parameters to identify the model. This point will be further addressed. In these equations, the parameters α_i and β_i are the random intercept and random slope for individual i; μ_α and μ_β are the mean of the intercepts and the mean of the slopes, respectively, and are fixed-effects parameters that are the same for all individuals; ϵ_{it}, ζ_{α_i}, and ζ_{β_i} are random errors; and λ_t is a time indicator that is a constant to which different values can be assigned to produce growth curve of different shapes that are linearly or nonlinearly dependent on time. For example, in the case of the linear LCM, λ_t equals $t-1$ for all t.

For the data in Table 2.1, which involves analysis of medical data with an initial measurement and after-treatment measurement of the test subjects, we have $T = 2$, λ_1=0, and $\lambda_2 = 1$. As a result, a linear growth curve for individual i can be obtained from (2.2), which is a straight line connecting the points $(0, \alpha_i + \epsilon_{i1})$ and $(1, \alpha_i + \beta_i + \epsilon_{i2})$ in the 2-dimensional plane (t, y_{it}) and is presented in Figure 2.1. The linear growth curve for individual i, $i = 1, \ldots N$ has intercept $\alpha_i + \epsilon_{i1}$ and slope $\beta_i + \epsilon_{i2} - \epsilon_{i1}$, and each individual has his/her own linear curve. However, statistical inference on the population growth curve that is determined by the mean of intercepts μ_α and the mean of slopes μ_β enables an examination of the existence of treatment effects.

Omitting individual index i, the general model (2.2) with $T = 2$, λ_1=0, and $\lambda_2 = 1$ can be expressed in matrix form:

$$Y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \Lambda\eta + \epsilon = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \tag{2.3}$$

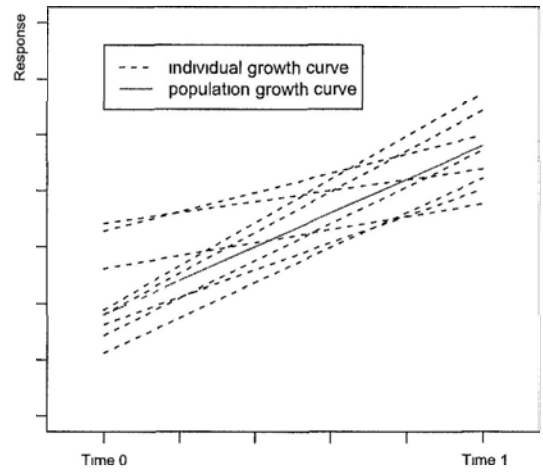

Figure 2.1: Growth curves

and

$$\eta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \mu_\eta + \zeta = \begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix} + \begin{pmatrix} \zeta_\alpha \\ \zeta_\beta \end{pmatrix}, \tag{2.4}$$

where Y is a 2×1 vector with measures y_1 and y_2 at the two time points, Λ is a 2×2 constant matrix of factor loadings, η is a 2×1 vector of two latent factors, ϵ is a 2×1 vector of residuals, μ_η is the vector of factor means, and ζ is a vector of residuals The two factors are, respectively, the intercept and slope of the growth curve. Equations (2.3) and (2.4) give the reduced form of Y:

$$Y = \Lambda(\mu_\eta + \zeta) + \epsilon. \tag{2.5}$$

With the commonly used assumptions that the distribution of ϵ is $N(0, \Theta_\epsilon)$ and that ζ is $N(0, \Psi)$, the mean and covariance matrix of Y are, respectively, given by

$$\mu_Y = \Lambda\mu_\eta \quad \text{and} \tag{2.6}$$

$$\Sigma_Y = \Lambda\Psi\Lambda' + \Theta_\epsilon. \tag{2.7}$$

For simplicity and clarity of presentation, we denote the two diagonal elements of Ψ that correspond to the intercept and the slope by $\psi_{\alpha\alpha}$ and $\psi_{\beta\beta}$, respectively, and the off-diagonal element by $\psi_{\alpha\beta}$.

Let the observations in a sample with size N for y_{i1}^* and y_{i2}^* be organized as frequencies in a $K \times K$ contingency table, and let $N_{k_1k_2}$ denote the number of cases that fall into the (k_1, k_2) cell of this contingency table. Let $\tau_k, k = 1, 2, \ldots, K-1$, be the thresholds for the variables y_1^* and y_2^*, and let θ be the vector that collects all of the unknown parameters, including the thresholds and the unknown parameters in μ_Y of (2.6) and Σ_Y of (2.7), then, the maximum likelihood estimates (MLE) of unknown parameters can be obtained by maximizing the following log-likelihood.

$$lnL = C + \Sigma_{k_1=1}^{K} \Sigma_{k_2=1}^{K} N_{k_1k_2} ln(\pi_{k_1k_2}(\theta)), \tag{2.8}$$

where C is a constant, and

$$\pi_{k_1k_2}(\theta) = P(y_1 = k_1, y_2 = k_2) = \int_{\tau_{k_1-1}}^{\tau_{k_1}} \int_{\tau_{k_2-1}}^{\tau_{k_2}} \phi_2(u, v; \mu_Y, \Sigma_Y) du dv, \tag{2.9}$$

where $\phi_2(u, v; \mu_Y, \Sigma_Y)$ is the density of the bivariate normal distribution with mean μ_Y and covariance matrix Σ_Y. It is worthy of note that, as ordinal categorical variables do not have an origin and unit of measurement, not all of the parameters in θ are identified, and hence constraints must be imposed to identify the model. We impose constraints to allow easy interpretation of the model parameters.

The MLE of the model parameters can be obtained by maximizing the log-likelihood. As the model in (2.5) takes the form of a factor analysis model, statistical software programs used to analyze factor analysis models can be used accordingly to produce the parameter estimates. For example, we use the Mx program (Neale *et al.*, 1999) to find the MLE. Mx is chosen because it is available in the public domain for free downloading, which enhances the accessibility of the proposed method to practitioners.

As the factor analysis model is a special case of the structural equation model (SEM), other methods that are available in the SEM literature for analyzing SEM

Mx Results				LISREL Results			
Par.	Est.	Std Err	t value	Par.	Est.	Std Err	t value
μ_α	0*	/	/	μ_α	0*	/	/
μ_β	-1.016	0.135	-7.550	μ_β	-0.990	0.071	-13.976
$\psi_{\alpha\alpha}$	1*	/	/	$\psi_{\alpha\alpha}$	1*	/	/
$\psi_{\beta\beta}$	1.012	0.201	5.044	$\psi_{\beta\beta}$	0.986	0.189	5.206
$\psi_{\alpha\beta}$	-0.626	0.096	-6.527	$\psi_{\alpha\beta}$	-0.628	0.091	-6.921

* fixed parameters

Table 2.2: Estimates of the model parameters for the active drug group

with ordinal categorical data can also be applied to produce the estimates of the model parameters. For example, a two-stage method can be implemented in the very popular SEM computer software packages PRELIS (Jöreskog and Sörbom, 1999) and LISREL (Jöreskog and Sörbom, 2004) to produce these estimates. The first stage involves the use of PRELIS to produce consistent estimates for the threshold parameters and the elements in μ_Y and Σ_Y, and the second stage involves the use of LISREL to produce weighted least squares estimates of the unknown parameters on the right hand sides of (2.6) and (2.7).

We used the proposed model to analyze the data on "Active drug" in Table 2.1, and we fix $\mu_\alpha = 0$ and $\psi_{\alpha\alpha} = 1$ to identify the model. The MLE produced by Mx and the two-stage estimates produced by PRELIS and LISREL are presented in Table 2.2, from which it can be seen that the two sets of estimates are very close. More specifically, the estimates for μ_β are both negative with large t-values. The results indicate that μ_β is significantly different from 0, and hence we can conclude that the drug has a significant effect. The estimates (standard errors) of the thresholds generated by Mx are -1.359 (0.158), -0.519 (0.112), and 0.229 (0.115), respectively. The estimates of the thresholds given by PRELIS are -1.277, -0.616, and 0.266, respectively. Note that PRELIS does not produce the the standard errors of the estimates of these thresholds.

2.3 Comparing Two Treatments

Using the LCM, statistical analysis can be easily and effectively conducted to compare the two different treatments. Let $Y^{*(R)} = (y_1^{*(R)}, y_2^{*(R)})'$ be the observed ordinal variables at the two time points for a reference treatment group, and $Y^{*(T)} = (y_1^{*(T)}, y_2^{*(T)})'$ be those for the new treatment group. Similar to (2.1), we assume that all four variables in $Y^{*(R)}$ and $Y^{*(T)}$ are related to the continuous variables in $Y^{(R)} = (y_1^{(R)}, y_2^{(R)})'$ and $Y^{(T)} = (y_1^{(T)}, y_2^{(T)})'$ via the threshold model for the same set of thresholds $-\infty = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_{K-1} < \tau_K = +\infty$. The assumption of equal thresholds is employed to facilitate ease of interpretation, and this assumption is a tenable one when the test subjects' responses are assessed by the same group of clinicians using the same set of criteria. To compare the two groups, two LCMs of the forms in (2.3) and (2.4) are fitted simultaneously. More specifically, for $g = R$, the reference group, and $g = T$, the treatment group, we have

$$Y^{(g)} = \begin{pmatrix} y_1^{(g)} \\ y_2^{(g)} \end{pmatrix} = \Lambda\eta^{(g)} + \epsilon^{(g)} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \alpha^{(g)} \\ \beta^{(g)} \end{pmatrix} + \begin{pmatrix} \epsilon_1^{(g)} \\ \epsilon_2^{(g)} \end{pmatrix} \tag{2.10}$$

and

$$\eta^{(g)} = \begin{pmatrix} \alpha^{(g)} \\ \beta^{(g)} \end{pmatrix} = \mu_\eta^{(g)} + \zeta^{(g)} = \begin{pmatrix} \mu_\alpha^{(g)} \\ \mu_\beta^{(g)} \end{pmatrix} + \begin{pmatrix} \zeta_\alpha^{(g)} \\ \zeta_\beta^{(g)} \end{pmatrix}. \tag{2.11}$$

The MLE of the model parameters can be obtained by maximizing a likelihood function that aggregates two components, each of the form in (2.8), and is contributed by a group. To impose identification constraints that facilitate treatment comparisons, we fix the origin and unit of measurement for $y_1^{(R)}$ (or equivalently $\alpha^{(R)}$) by assuming that $y_1^{(R)}$ has mean zero and variance 1. In other words, it is assumed that $\mu_\alpha^{(R)} = 0$ and the first diagonal element, $\psi_{\alpha\alpha}^{(R)}$ of $\Psi^{(R)}$, equals 1. With these constraints and the assumption of the same thresholds across all variables, the model is identified, and all of the parameters are estimable.

We used the Mx program (Neale *et al.*, 1999) to implement the analysis. The Mx program code used to analyze the data in Table 2.1 is given in Appendix A-1, and

the results are presented in Table 2.3. The thresholds, and hence their estimates, are assumed to be the same for all four of the variables. We therefore have three threshold parameters. The estimates (standard errors) of the three parameters are -1.319 (0.152), -0.563 (0.109), and 0.242 (0.115), respectively. The value of the maximum log-likelihood is given by $L_0 = -18.002$.

We also used PRELIS and LISREL to produce the two-stage estimates for the model parameters, and the results are also shown in Table 2.3. They are close to those produced by Mx. The estimates of the thresholds are -1.277, -0.616, and 0.266, respectively.

From the results in Table 2.3, we can draw similar conclusions from the estimation results produced by Mx and LISREL. For example, all parameters are statistically significant except the one for the mean of the intercept of the treatment group. The t-value for $\mu_\alpha^{(T)}$ that are produced by Mx and LISREL are 0.291 and 0.494 respectively. The results indicate that the initial conditions of the subjects in the two treatments have no significant difference. This is consistent with what is expected because subjects have been randomly allocated to the two treatments.

A graphical representation of the estimated latent growth curves will also assist in the interpretation. Figure 2.2 presents the two estimated latent growth curves based on the Mx results. The intercepts of the two curves are close ($\mu_\alpha^{(R)} = 0^*$, $\mu_\alpha^{(T)} = 0.046$), thus suggesting that there is little difference between the initial conditions of the subjects in the two groups. The estimates of the slopes in both groups are negative ($\mu_\beta^{(R)} = -1.003$, $\mu_\beta^{(T)} = -0.645$), which suggests that both treatments will result in the subjects taking less time to fall asleep. The magnitude of the linear curve slope for the placebo group is smaller, which indicates that the downward trend in this group is smaller than that in the treatment group.

Mx Results				LISREL Results			
Active Drug (Reference Group)				Active Drug (Reference Group)			
Par	Est	Std Err	t value	Par	Est	Std Err	t value
$\mu_\alpha^{(R)}$	0*	/	/	$\mu_\alpha^{(R)}$	0*	/	/
$\mu_\beta^{(R)}$	-1 003	0 132	-7 597	$\mu_\beta^{(R)}$	-0 990	0 083	-11 941
$\psi_{\alpha\alpha}^{(R)}$	1*	/	/	$\psi_{\alpha\alpha}^{(R)}$	1*	/	/
$\psi_{\beta\beta}^{(R)}$	0 986	0 188	5 234	$\psi_{\beta\beta}^{(R)}$	1 129	0 226	4 995
$\psi_{\alpha\beta}^{(R)}$	-0 638	0 092	-6 918	$\psi_{\alpha\beta}^{(R)}$	-0 565	0 106	-5 316
	Placebo (Treatment group)				Placebo (Treatment group)		
$\mu_\alpha^{(T)}$	0 046	0 158	0 291	$\mu_\alpha^{(T)}$	0 045	0 092	0 494
$\mu_\beta^{(T)}$	-0 645	0 129	-4 986	$\mu_\beta^{(T)}$	-0 645	0 079	-8 180
$\psi_{\alpha\alpha}^{(T)}$	1 228	0 375	3 278	$\psi_{\alpha\alpha}^{(T)}$	1 000	0 092	10 909
$\psi_{\beta\beta}^{(T)}$	0 895	0 264	3 391	$\psi_{\beta\beta}^{(T)}$	0 741	0 201	3 694
$\psi_{\alpha\beta}^{(T)}$	-0 485	0 221	-2 194	$\psi_{\alpha\beta}^{(T)}$	-0 370	0 119	-3 102

* fixed parameters

Table 2 3 Maximum likelihood estimates of the model parameters

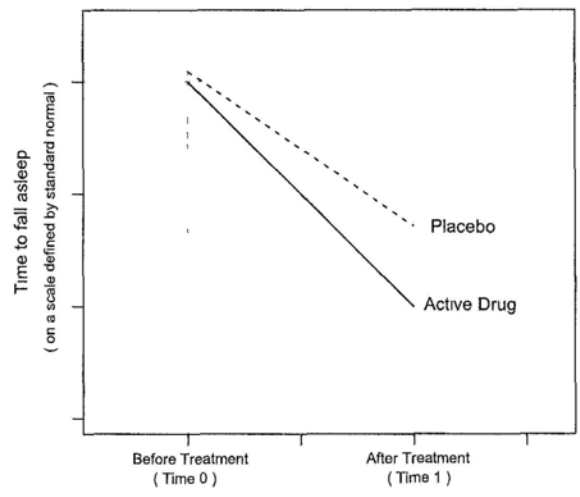

Figure 2.2: The latent growth curve for the active drug and placebo treatments

Different hypotheses on the intercepts and slopes of the two linear curves can be tested to assess whether various effects are significant. Several typical tests are as follows

Test 1: $H_0^{(1)} : \mu_\alpha^{(T)} = 0$ *vs.* $H_1^{(1)} : \mu_\alpha^{(T)} \neq 0$.

This test is used to examine the means of the intercepts of the two groups, that is, to examine whether the initial conditions of these groups are the same. From the Mx results in Table 2.3, we know that the estimated value of $\mu_\alpha^{(T)}$ is 0.046 and not statistically significant. A likelihood ratio test can also be constructed to test the hypothesis. By imposing the constraint $\mu_\alpha^{(T)} = 0$ and maximizing the log-likelihood, the value of the maximum log-likelihood is given by $L_1 = -18.045$. As $-2(L_1 - L_0) = 0.086$, comparing the value to the chi-squared distribution with 1 degree of freedom will reproduce the aforementioned conclusion that the initial conditions of the two groups, in terms of the means of the intercepts, are the same.

Test 2: $H_0^{(2)} : \mu_\alpha^{(T)} = 0$ *and* $\psi_{\alpha\alpha}^{(T)} = 1$ *vs.* $H_1^{(2)}$*: at least one of the equalities does not*

hold.

This test is used to examine whether the initial conditions of the two groups, in terms of the means and variances of the intercepts, are the same. A likelihood ratio test can be used. The value of the maximum log-likelihood with the constraints $\mu_\alpha^{(T)} = 0$ and $\psi_{\alpha\alpha}^{(T)} = 1$ is given by $L_2 = -18.233$. As $-2(L_2 - L_0) = 0.462$, comparing it to the chi-squared distribution with 2 degrees of freedom does not lead to the rejection of the null hypothesis, thus suggesting that the initial conditions of the two groups are the same.

Test 3: $H_0^{(3)} : \mu_\beta^{(R)} = \mu_\beta^{(T)}$ *vs.* $H_1^{(3)} : \mu_\beta^{(R)} \neq \mu_\beta^{(T)}$.

This test examines whether there is any treatment difference by testing whether the means of the slopes of the growth curves are the same in the two groups. A likelihood ratio test can again be used. The value of the maximum log-likelihood with the constraint $\mu_\beta^{(R)} = \mu_\beta^{(T)}$ is given by $L_3 = -20.585$. As $-2(L_3 - L_0) = 5.165$, comparing it to the chi-squared distribution with 1 degree of freedom suggests that the means of the slopes of the growth curves for both groups are significantly different. In other words, the effects of the two treatments are different.

Test 4: $H_0^{(4)} : \mu_\beta^{(R)} = \mu_\beta^{(T)}$ *and* $\psi_{\beta\beta}^{(R)} = \psi_{\beta\beta}^{(T)}$ *vs.* $H_1^{(4)}$*: at least one of the equalities does not hold.*

This test examines whether there is any treatment difference by testing whether both the means and the variances of the slopes of the growth curves are the same for both groups. $H_0^{(4)}$ can be regarded as a more stringent hypothesis than $H_0^{(3)}$ because it requires not only the locations of these slopes, but also the dispersions of their distributions, to be the same in the two treatments. Again, a likelihood ratio test can be used. The value of the maximum log-likelihood with the constraints $H_0^{(4)} : \mu_\beta^{(R)} = \mu_\beta^{(T)}$ and $\psi_{\beta\beta}^{(R)} = \psi_{\beta\beta}^{(T)}$ is given by $L_4 = -20.597$. As $-2(L_4 - L_0) = 5.190$, comparing it to the chi-squared distribution with 2 degrees of freedom provides a p-value of 0.075, thus suggesting that the result is marginal. With a Type-I error of 0.1, either the mean or the variance of the slope of the growth curve for the reference group, or both, is significantly different from that of the treatment group. In other

words, the two treatments have different effects. However, with a Type-I error of 0.05, we cannot conclude that there is a significant treatment difference.

It is worthy of note that as the slope of the growth curve represents the rate of change, Tests 3 and 4, which use the slopes of the two treatment groups as the basis for detecting a treatment difference, have the nice feature that their results are valid no matter the initial conditions (i.e., the intercepts) of the two groups are the same or not.

Other types of tests that serve different purposes can also be constructed. For example, one may be interested in testing whether the variances of the slopes of the two groups are the same. Moreover, if the hypothesis $H_0^{(1)}$ or $H_0^{(2)}$ is not rejected, that is, when the initial conditions of the two groups can be regarded as the same, then a more restricted model with equal initial conditions for the two groups can be used as the alternative model to achieve greater power in testing for the presence of treatment difference. Table 2.4 provides a summary of the aforementioned tests. The various test results in this table have provided clear answers to the first three research questions (Q1 to Q3, Section 2.1). It leads to the conclusion that the initial conditions of the subjects in the two treatment groups are the same and the effects of the two treatments are different. That is, over the two-week treatment period, there is, on average, a more significant downward trend in the time taken to fall asleep in the active drug group than in the placebo group. These results are sensible. As the patients with insomnia were randomly assigned to the active hypnotic drug group and the placebo group, the initial conditions of the subjects in these groups should not be different. It is worthy of note that, although the conclusion of the existence of a treatment difference is only marginal in Test 4, it becomes clear in Test 8 that this difference is significant. These results demonstrate that by making use of the finding that the initial conditions of the two treatment groups are the same to formulate a more restricted model in the alternative hypothesis, it becomes possible to achieve a test with greater power. This example also demonstrates that the use of the LCM allows various hypotheses with simple and straightforward interpretations

Model i/ Test i	H_0	H_1 Model	L_i under H_0	Test Statistic	Test Result	p-value
0	—	—	$L_0 = -18.002$	—	—	—
1	$\mu_\alpha^{(T)} = 0$	Model 0	$L_1 = -18.045$	$-2(L_1 - L_0)$	0.086	0.769
2	$\mu_\alpha^{(T)} = 0$ $\psi_{\alpha\alpha}^{(T)} = 1$	Model 0	$L_2 = -18.233$	$-2(L_2 - L_0)$	0.462	0.794
3	$\mu_\beta^{(R)} = \mu_\beta^{(T)}$	Model 0	$L_3 = -20.585$	$-2(L_3 - L_0)$	5.165	0.023
4	$\mu_\beta^{(R)} = \mu_\beta^{(T)}$ $\psi_{\beta\beta}^{(R)} = \psi_{\beta\beta}^{(T)}$	Model 0	$L_4 = -20.597$	$-2(L_4 - L_0)$	5.190	0.075
5	$\mu_\beta^{(R)} = \mu_\beta^{(T)}$	Model 1	$L_5 = -22.956$	$-2(L_5 - L_1)$	9.822	0.0017
6	$\mu_\beta^{(R)} = \mu_\beta^{(T)}$	Model 2	$L_6 = -23.487$	$-2(L_6 - L_2)$	10.508	0.0012
7	$\mu_\beta^{(R)} = \mu_\beta^{(T)}$ $\psi_{\beta\beta}^{(R)} = \psi_{\beta\beta}^{(T)}$	Model 1	$L_7 = -22.962$	$-2(L_7 - L_1)$	9.833	0.0073
8	$\mu_\beta^{(R)} = \mu_\beta^{(T)}$ $\psi_{\beta\beta}^{(R)} = \psi_{\beta\beta}^{(T)}$	Model 2	$L_8 = -23.548$	$-2(L_8 - L_2)$	10.629	0.0049

Table 2.4: Summary of tests

to be set up and tested in a systematic efficient and easy manner.

2.4 Other Applications of the LCM

Section 2.3 summarizes the way in which the LCM can be used to compare two treatment groups. In effect, the use of the LCM to model and analyze the type of data presented in Table 2.1 facilitates the further examination of many topics that are relevant to treatment comparisons. We explore two major topics in this section: Analysis of covariate effects and statistical inference for equivalent treatments.

2.4.1 Analyzing the Effects of Covariates

Analyzing the effects of covariates is important in many medical studies. For example, gender, age, race, etc. are usually included in an analysis as covariates to further

explain the response variables. We further generalize the LCM model in (2.2) to incorporate time-invariant covariates or explanatory variables. We consider a model in which the covariates have a direct influence on the random intercepts and slopes. These covariates may be dummy-coded categorical variables or measured variables on an interval scale. More specifically, an LCM with time-invariant covariates can be obtained by generalizing the model in (2.2). It can be expressed as follows

$$\begin{aligned} \text{Trajectory equation:} \quad & y_{it} = \alpha_i + \lambda_t \beta_i + \epsilon_{it} \\ \text{Intercept equation:} \quad & \alpha_i = \mu_\alpha + \gamma_{\alpha 1} x_{1i} + \cdots + \gamma_{\alpha k} x_{ki} + \zeta_{\alpha i} \qquad (2.12) \\ \text{Slope equation:} \quad & \beta_i = \mu_\beta + \gamma_{\beta 1} x_{1i} + \cdots + \gamma_{\beta k} x_{ki} + \zeta_{\beta i} \qquad (2.13) \end{aligned}$$

In these equations, $x_{1i}, x_{2i}, \ldots, x_{ki}$ are covariates that are independent of $\zeta_{\alpha i}$ and $\zeta_{\beta i}$. Similar to (2.3) and (2.4), the matrix form for this model can be written as

$$Y = \Lambda \eta + \epsilon \qquad (2.14)$$

$$\eta = \mu_\eta + \Gamma X + \zeta, \qquad (2.15)$$

where

$$\Gamma = \begin{pmatrix} \gamma_{\alpha 1} & \gamma_{\alpha 2} & \cdots & \gamma_{\alpha k} \\ \gamma_{\beta 1} & \gamma_{\beta 2} & \cdots & \gamma_{\beta k} \end{pmatrix}$$

$$X = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ki} \end{pmatrix}.$$

The reduced-form model is given by

$$Y = \Lambda(\mu_\eta + \Gamma X) + \Lambda \zeta + \epsilon, \qquad (2.16)$$

and the mean and covariance matrix of Y are, respectively, given by

$$\mu_Y = \Lambda(\mu_\eta + \Gamma \mu_X), \quad \text{and} \qquad (2.17)$$

$$\Sigma_Y = \Lambda(\Gamma \Sigma_X \Gamma' + \Psi)\Lambda' + \Theta_\epsilon. \qquad (2.18)$$

From (2.16), we have $cov(Y, X) = \Lambda\Gamma\Sigma_X$, and hence the mean and covariance matrix of all of the observed covariates and the underlying continuous variables are given by

$$\mu = \begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix} = \begin{pmatrix} \Lambda(\mu_\eta + \Gamma\mu_X) \\ \mu_X \end{pmatrix}, \quad \text{and} \tag{2.19}$$

$$\Sigma = cov \begin{pmatrix} Y \\ X \end{pmatrix} = \begin{pmatrix} \Lambda(\Gamma\Sigma_X\Gamma' + \Psi)\Lambda' + \Theta_\epsilon & \Lambda\Gamma\Sigma_X \\ \Sigma_X\Gamma'\Lambda' & \Sigma_X \end{pmatrix}. \tag{2.20}$$

Estimates of the model parameters can be implemented in popular SEM software packages with appropriately designed programs codes.

As an illustration, we analyze the data in Table 2.1. We introduce a dummy variable that represents the two treatment groups into the LCM as a time-invariant covariate to test the possible difference between the two groups. This model is given by

$$y_{it} = \alpha_i + \lambda_t\beta_i + \epsilon_{it}, \tag{2.21}$$

$$\alpha_i = \mu_\alpha + \gamma_\alpha X_i + \zeta_{\alpha i}, \quad \text{and} \tag{2.22}$$

$$\beta_i = \mu_\beta + \gamma_\beta X_i + \zeta_{\beta i}, \tag{2.23}$$

where X_i is a dummy variable that takes a value of 1 if subject i is in the active drug group and of 0 if it is in the placebo group. Based on a simple statistical test, we can check whether there exists any significant difference between the groups.

We also used the Mx program to find the MLEs of the parameters in the LCM with covariates. The Mx input script can be found in Appendix A-2. It should be noted that, for the Mx data input, we transformed the data in Table 2.1 into the values of two ordinal categorical variables, say Y_1 and Y_2, where Y_1 denotes the ordinal categorical response at time 1 with categories 0, 1, 2 and 3, and Y_2 is the ordinal categorical response at time 2 with the same categories. A new dummy variable, say X, is added as the third variable in the ordinal data file. The results are reported in Table 2.5.

Mx Results				LISREL Results			
Par.	Est.	Std Err	t value	Par.	Est.	Std Err	t value
μ_α	0*	/	/	μ_α	0*	/	/
μ_β	-0.558	0.114	-4.879	μ_β	-0.785	0.052	-15.133
γ_α	-0.037	0.178	-0.211	γ_α	-0.008	0.089	-0.087
γ_β	-0.465	0.181	-2.568	γ_β	-0.268	0.092	-2.912
$\psi_{\alpha\alpha}$	1*	/	/	$\psi_{\alpha\alpha}$	1*	/	/
$\psi_{\beta\beta}$	0.831	0.124	6.674	$\psi_{\beta\beta}$	0.859	0.150	5.715
$\psi_{\alpha\beta}$	-0.513	0.074	-6.960	$\psi_{\alpha\beta}$	-0.468	0.071	-6.541

* fixed parameters

Table 2.5: Parameters estimates in the LCM with a dummy variable covariate

Using the results in Table 2.5, we can investigate the differences between the two groups by studying the magnitude and significance of the coefficient of the dummy variable X, which takes a value of 0 for the placebo group and of 1 for the drug group. First, for the latent intercepts, the difference between the placebo group and the active drug group is very small ($\gamma_\alpha = -0.037$) and is not statistically significant (the t value is -0.211), thus suggesting that the initial conditions of the subjects in the two groups are the same. For the latent slopes, the difference between the two groups is -0.465 and statistically significant (the t value is -2.568), which suggests a significant treatment difference. There is also a significant treatment effect for the placebo group ($\mu_\beta = -0.558$ with t value = -4.879), even though the treatment is a placebo. This significant placebo effect suggests that it is also worth exploring the possibility of establishing equivalence between the placebo and the active drug, which we do in the following section.

The PRELIS and LISREL two-stage results are also reported in Table 2.5. The conclusions based on the two-stage estimates are consistent with those based on the Mx MLE estimates.

In this subsection, we introduced the method to incorporate time-invariant covariates into the LCM. As a typical case, by incorporating a dummy variable indicating different treatment into LCM as covariate, we can conveniently draw inferences

about whether there are significant differences between two treatments. For the original data presented in Table 2.1 of the example, we draw completely consistent results when we use this method and the method described in section 2.3. Specifically, the active drug and placebo have the same initial conditions; after a two-week treatment, the active drug shows significant treatment difference with the placebo. By using appropriate dummy variables, the method can be applied to compare more than two groups.

2.4.2 Inference for Equivalence Treatments

The tests discussed in Section 2.3, which are used to compare two treatments, have focused on examining whether the values of the parameters that represent the two treatment groups are the same. In many medical research studies, such criteria are found to be too stringent. For example, when there is a new drug that is less expensive than or has fewer side effects than an existing drug, consensus has been reached in the medical field that the primary objective is not to show that the two drugs are equally effective, but to show that the new drug is equivalent to a standard drug. That is, the difference between the parameters that characterize the two treatments is less than a small, pre-defined margin. Simply speaking, for a given Δ, if the effectiveness of a standard reference drug and a new treatment are represented by two continuous random variables with location parameters $\mu^{(R)}$ and $\mu^{(T)}$, then the equivalence of these two drugs, in terms of the difference in their locations, is assessed by testing the following hypothesis.

$$H_0 : \mu^{(T)} - \mu^{(R)} \leq -\Delta \quad \text{or} \quad \mu^{(T)} - \mu^{(R)} \geq \Delta \quad \text{vs.} \quad H_1 : -\Delta < \mu^{(T)} - \mu^{(R)} < \Delta. \tag{2.24}$$

Equivalence can be concluded at Type-I error level α for a given Δ value if H_0 is rejected, that is, when the $100 \times (1-2\alpha)\%$ confidence interval $[D_L, D_U]$ for $\mu^{(T)} - \mu^{(R)}$ falls entirely within $[-\Delta, \Delta]$.

In the literature on equivalent tests, the ratio of the means has also been widely used in assessing the equivalence of two treatments. For a given δ, the equivalence of two drugs in terms of this ratio is assessed by testing the following hypothesis.

$$H_0 : \mu^{(T)}/\mu^{(R)} \leq \delta \quad \text{or} \quad \mu^{(T)}/\mu^{(R)} \geq 1/\delta \quad \text{vs.} \quad H_1 : \delta < \mu^{(T)}/\mu^{(R)} < 1/\delta. \quad (2.25)$$

In this case, equivalence can be concluded at Type-I error level α for a given δ value if H_0 is rejected, that is, when the $100 \times (1 - 2\alpha)\%$ confidence interval $[R_L, R_U]$ for $\mu^{(T)}/\mu^{(R)}$ falls entirely within $[\delta, 1/\delta]$.

When the LCM is employed to analyze ordinal response data, hypotheses in the forms of (2.24) and (2.25) can easily be constructed based on the means of the slopes of the LCMs to examine the equivalence of two treatments. More specifically, we use a hypothesis in the form of (2.25) to examine the equivalence of the slopes of the growth curves. As an illustration, we seek an answer for the research question Q4 (see Section 2.1) by examining whether the placebo is equivalent to the active drug for the data set presented in Table 2.1 by testing the following hypothesis

$$H_0 : \mu_\beta^{(T)}/\mu_\beta^{(R)} \leq \delta \quad \text{or} \quad \mu_\beta^{(T)}/\mu_\beta^{(R)} \geq 1/\delta \quad \text{vs.} \quad H_1 : \delta < \mu_\beta^{(T)}/\mu_\beta^{(R)} < 1/\delta. \quad (2.26)$$

Following the common practice (see, e.g. Tang and Poon, 2007), we set $\delta = 0.8$. The 95% confidence interval [0.388, 0.898] for $\mu_\beta^{(T)}/\mu_\beta^{(R)}$ for the data set in Table 2.1 does not fall entirely into the interval $[\delta, 1/\delta] = [0.8, 1.25]$. In other words, the hypothesis that the placebo is equivalent to the active drug is rejected.

As the tests for the same initial conditions (Tests 1 and 2 in Table 2.4) are not rejected, we again find the 95% confidence interval for $\mu_\beta^{(T)}/\mu_\beta^{(R)}$ under the constraints $H_0^{(2)} : \mu_\alpha^{(T)} = 0$ and $\psi_{\alpha\alpha}^{(T)} = 1$, which is the null model in Test 2. As the resultant interval [0.418, 0.807] does not fall entirely into the interval $[\delta, 1/\delta] = [0.8, 1.25]$, we come to the same conclusion and reject the hypothesis that the placebo is equivalent to the active drug.

2.5 Conclusion

In this chapter, we propose the use of a latent growth curve to model and analyze ordinal categorical data that involve measurements at two different time points. There are several prominent advantages to using such an approach. First, the model has easy and straightforward interpretations and can be represented graphically, thus enhancing its accessibility to practitioners. Second, it can be generalized to conduct many different types of analyses that are important in medical and other studies. For example, we have discussed how the model can be used to compare treatment effects, to incorporate and analyze covariates, and to establish treatment equivalence. Many other generalizations are also possible. Third, as the initial conditions and the effects of the treatments are represented by the intercept and slope, respectively, in a comparison of two treatments or in the establishment of their equivalence, statistical tests based on intercepts and slopes can be constructed accordingly to examine the possible difference between initial conditions and to test the significance of the treatment effect. The results derived from testing the treatment effect remain valid, even though the initial conditions are not the same. Moreover, if the same initial conditions can be established for the participating subjects, then tests with greater power can easily be formulated by using a more restricted alternative model. Fourth, the latent growth curve model is a special case of the SEM, and hence, with appropriate specifications, the estimates of the model parameters can be obtained using a variety of widely accessible SEM computer programs. We have provided a sample program for implementing the procedure in Mx, which can be downloaded in the public domain, thus enhancing the accessibility of the proposed approach. Finally, many generalizations of this model are possible. The availability of numerous software programs, and their continuous development, further enhances the accessibility of the newly developed methods.

Our discussion has focused on analysis of a data set that involves measurements of two different time points. When a study involves measurements at more than two

different time points, the model can easily be generalized to analyze the longitudinal data of a study that involves measurements at more than two time points. In this case, not only the linear, but also the non-linear growth curve can be analyzed.

We have discussed the comparison of two treatments. The generalization of the method to a comparison of three or more treatments represents an interesting topic for further study.

The approach of the analysis of variance and covariance models is commonly used for treatment comparisons. These models are in general applicable to variables that are continuous in scale, and cannot be directly applied to analyze ordinal categorical data. As well established statistical methods can be used to analyze the class of analysis of variance and covariance models, how these models can be generalized to analyze ordinal categorical variables represents an interesting topic for further research. The use of covariates in these models will also facilitate the analysis of initial conditions.

Chapter 3

Multiple Testing of Several Treatments with a Control

3.1 Introduction

In the previous chapter, we considered the modeling and analysis of longitudinal ordinal responses that involve measurements at two different time points. By modeling the longitudinal ordinal responses using the latent growth curve, not only the initial conditions of the two treatments but also the trend and extent of the treatment effects can be compared. In this chapter, we will consider another very common treatment comparison issue: the comparison of several treatments with a control.

For the comparison of several treatments with a control, one important task is the control of the family-wise error (FWE) at preassigned value α, which is defined as

$$\text{FWE} = \text{P(reject any true hypothesis)} \leq \alpha.$$

Another important task is to improve the power of the testing procedure. In the literature of multiple comparison, there are several possible definitions of power (see Horn and Dunnett, 2004), such as any-pair power, all-pairs power, and average power. Let m be the number of false hypotheses in a multiple testing, and Z_m be the random number of rejected false hypotheses. Then, the several different power can be defined as

- any-pair power: P_{any}=P($Z_m \geq 1$)
- all-pairs power: P_{all}=P($Z_m = m$)
- average power: P_{ave}=$E(Z_m)/m = \sum_{t=1}^{m} P(Z_m \geq t)/m$

Many multiple testing procedures, both single-step and stepwise procedures, have been developed. Some of them will be detailed discussed in Section 3.5.2.

In this chapter, we focus on the comparisons of several treatments to a control that have ordered categorical responses. To compare two treatments with ordered categorical responses, a very popular approach is to adopt the logistic regression model with the proportional odds assumption (McCullagh, 1980). Then, the Wilcoxon-Mann-Whitney (WMW) test (Wilcoxon, 1945; Mann and Whitney, 1947) can be applied to test for equality of treatment efficacy. Score statistics can be used for the WMW test (Whitehead, 1993). However, our study (see section 3.3) finds that this existing method can not preserve the type I error when the two treatments have different dispersions. This motivates us to find new testing method that can accommodate the difference in the dispersions of the treatments.

In this chapter, we will develop new testing method for the comparison of several treatments with a control that have ordinal responses. Our interest lies in comparing the mean efficacy (the location parameter) of treatments, that is, in exploring whether a treatment is better than the control on average, amid the possibility of having different variances among the treatment groups. Our analysis is based on the latent variable model, that is, the ordinal responses are regarded as the manifestations of some underlying continuous random variables. Within this framework, we will demonstrate that the proportional odds model approach may yield too large a probability of rejecting a true null hypothesis that tests the equality of treatment means when the proportional odds assumption is invalid. A two-step estimation procedure is proposed for the parameter estimation of the latent variable model in Section 3.4. Multiple comparison procedures are proposed in Section 3.5, including both single-step and stepwise procedures. The evaluation of power by a simulation

study is also offered in Section 3.5. Clinical examples to illustrate the implementation of our procedures are given in Section 3.6. Some technical proofs are given in the Appendix B.

3.2 The Latent Variable Model

Suppose that there are $G+1$ treatments, and n_i $(i = 0, 1, \ldots, G)$ patients receive treatment i where subscript 0 $(i = 0)$ denotes the control treatment. The responses of the subjects receiving treatments are classified into one of the categories $C_1, \cdots, C_K$. The ordinal responses for treatment i are considered to be manifestations of a continuous latent variable X_i which has a normal distribution with mean μ_i and variance σ_i^2, $i = 0, \ldots, G$. The use of latent variables provides a useful framework in which to model categorical responses. Many studies have used latent variables in the social sciences (see Lee, 2007, and references therein). However, the application of latent variables to compare treatments in clinical studies with ordered categorical data is less developed. To model ordinal categorical responses, Anderson and Philips (1981) incorporated the notion of latent variables into their formulation of the logistic model for prediction and discrimination, mainly using the variables to assist interpretation. Latent variable models have also been applied to analyzing clustered ordinal data (see Qu, Piedmonte, and Medendorp, 1995, and references therein). Bekele and Thall (2004) used latent variables and the Bayesian approach to model toxic responses in their dose-finding study in a phase I trial of gemcitabine for the treatment of soft tissue sarcoma. Their method was more recently extended to model ordinal data nested within categories (Leon-Novelo *et al.*, 2010). A common feature of the aforementioned models is that the variances of the treatments are either assumed to be identical or fixed at specified ratios.

Now, we outline our proposed latent variable model, which is more flexible and allows for heterogenous variances across different treatments. We assume that the ordinal categorical variable falls into category C_k if and only if the latent variable

	Categories				
Treatment	C_1	C_2	$\cdots$	C_K	Total
Control	n_{01}	n_{02}	$\cdots$	n_{0K}	n_0
Treatment 1	n_{11}	n_{12}	$\cdots$	n_{1K}	n_1
.	.	.	$\cdots$	.	.
.	.	.	$\cdots$	.	.
.	.	.	$\cdots$	.	.
Treatment G	n_{G1}	n_{G2}	$\cdots$	n_{GK}	n_G

Table 3.1: Ordered categorical data of a clinical trial.

satisfies $\tau_{k-1} < X_i \leq \tau_k$. For the convenience of comparing treatments, the thresholds τ_k, $k = 1, \ldots, K-1$, are assumed to be the same across all treatments, where $\tau_0 = -\infty$ and $\tau_K = \infty$. This assumption is also reasonable in practice because the subjects of different treatments are usually evaluated according to the same criterion and measured with the same device.

Let n_{ik} be the number of patients receiving treatment i, with outcomes classified into category C_k. Furthermore, let $P(\tau_{k-1} < X_i \leq \tau_k)$ be π_{ik}, the probability of a patient's outcome falling into category C_k with treatment i. Then, with respect to treatment i, we have

$$\sum_{k=1}^{K} n_{ik} = n_i, \quad \sum_{k=1}^{K} \pi_{ik} = 1, \quad i = 0, \ldots, G.$$

In addition, let $N = \sum_{i=0}^{G} n_i$ be the total sample size. The ordered categorical data at the end of a clinical trial can be summarized in Table 3.1.

The log-likelihood function for the samples of the $G+1$ treatments is given by

$$L(\boldsymbol{\theta}) = \sum_{i=0}^{G} \sum_{k=1}^{K} n_{ik} \log(\pi_{ik}(\boldsymbol{\theta})) \tag{3.1}$$

where

$$\begin{aligned} \pi_{ik}(\boldsymbol{\theta}) &= \int_{\tau_{k-1}}^{\tau_k} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu_i)^2}{\sigma_i^2}\right) dx \\ &= \Phi\left(\frac{\tau_k - \mu_i}{\sigma_i}\right) - \Phi\left(\frac{\tau_{k-1} - \mu_i}{\sigma_i}\right), \end{aligned} \tag{3.2}$$

$\Phi(\cdot)$ is the c.d.f. of the standard normal distribution and $\boldsymbol{\theta} = (\boldsymbol{\tau}', \boldsymbol{\theta}_0')'$ is a vector containing all of the unknown parameters, with $\boldsymbol{\tau}' = (\tau_1, \cdots, \tau_{K-1})$ and $\boldsymbol{\theta}_0' = (\mu_0, \cdots, \mu_G, \sigma_0, \cdots, \sigma_G)$.

Denote $\eta_i = \mu_i - \mu_0$ for $i = 1, \ldots, G$ as the difference in treatment efficacy between the treatment i and the control, $i = 1, \cdots, G$. Without loss of generality, assume that a positive value of η_i implies that treatment i is more effective. We are interested in testing the null hypothesis

$$H_i : \eta_i = 0 \tag{3.3}$$

against the two-sided alternative

$$H_i' : \eta_i \neq 0 \tag{3.4}$$

for $i = 1, ..., G$ simultaneously. As one-sided tests are straightforward generalizations of two-sided tests, we consider only two-sided tests for simplicity.

3.3 The Proportional Odds Model

The proportional odds model is frequently employed in clinical studies. For example, Diem *et al.* (2006) used it to study the effects of ultralow-dose transdermal estradiol on postmenopausal symptoms in women aged 60 to 80 years. In another study that evaluated immunogenicity in vaccine trials, Pédrono *et al.* (2009) recommended the proportional odds model based on clinical relevance and statistical power.

Now, we outline the proportional odds model and the testing procedure that compares one treatment to a control ($G = 1$). Let $\gamma_{ik} = \pi_{i1} + \cdots + \pi_{ik}$ be the cumulative probabilities for treatment i, $i = 0, 1$; $k = 1, \cdots, K$. Then, the log-odds ratios can be defined by

$$\gamma_k = \log \left\{ \frac{\gamma_{1k}(1 - \gamma_{0k})}{\gamma_{0k}(1 - \gamma_{1k})} \right\}, \quad k = 1, \cdots, K - 1. \tag{3.5}$$

The proportional odds model assumes that $\gamma_1 = \gamma_2 = \cdots = \gamma_{K-1} = \gamma$ where γ is the common odds. To test whether the treatment and the control have the same effects,

the null hypothesis is $H_0 : \gamma = 0$. Here, we follow the testing procedure given by Whitehead (1993). The efficient score statistic is

$$Z = \frac{1}{N+1} \sum_{k=1}^{K} n_{1k}(L_{0k} - U_{0k}), \tag{3.6}$$

where

$$L_{0k} = n_{01} + \cdots + n_{0(k-1)}, \qquad \text{for } k = 2, \cdots, K,$$

and

$$U_{0k} = n_{0(k+1)} + \cdots + n_{0K}, \qquad \text{for } k = 1, \cdots, K-1,$$

are the lower and upper cumulative responses of the control, respectively. In addition, take $L_{01} = U_{0K} = 0$. To test the null hypothesis, the test statistic is $Z/\sqrt{V}$, where

$$V = \frac{n_0 n_1 N}{3(N+1)^2} \left[1 - \sum_{k=1}^{K} \left(\frac{n_{0k} + n_{1k}}{N} \right)^3 \right]. \tag{3.7}$$

For a given significance level α, the null hypothesis will be rejected if

$$\frac{|Z|}{\sqrt{V}} > z_{\alpha/2}, \tag{3.8}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ probability point of the standard normal distribution.

The remainder of this section demonstrates an undesirable consequence of the above test when the proportional odds assumption fails to hold. As indicated by Peterson and Harrell (1990), clinical examples of non-proportional odds are not difficult to find, and they used a coronary artery disease data set to illustrate that the proportional odds model is not appropriate. Dark, Bolland and Whitehead (2003) also discussed the problem of using the proportional odds model for clinical trials where the validity of the proportional odds assumption is questionable. In fact, the popular χ^2 test for the proportional odds assumption is quite conservative. Dark, Bolland and Whitehead (2003) used the numerical example extracted from the National Institute of Neurological Disorders and Stroke t-PAstroke study to consider the proportional odds assumption. The estimated odds ratios computed from the data do not support the use of the proportional odds model even though the p-value is 0.134 for the χ^2 test.

Now, let us refer to the latent variable model given in Section 3.2. Without loss of generality, take $G = 1$. Consider the ordinal responses of the two treatments, denoted by Y_0 and Y_1, as manifestations of some underlying continuous normal variables, denoted by $X_0 \sim N(\mu_0, \sigma_0^2)$ and $X_1 \sim N(\mu_1, \sigma_1^2)$, respectively.

Nonhomogeneous treatment groups exert a significant influence on the WMW procedure. To illustrate, consider the following simple case with equal mean efficacy for the treatment and the control such that $X_0 \sim N(\mu, \sigma_0^2)$ and $X_1 \sim N(\mu, \sigma_1^2)$. Furthermore, let $1 < s < K$ where $\tau_{s-1} < \mu < \tau_s$. For simplicity, assume equal sample size for the two treatments, denoted by n. Hence, the total sample size is $N = 2n$. Now, assume that $\sigma_1^2 \to 0$. Therefore, the observations Y_1 will fall in the category C_s with probability approaching to 1. Then, with respect to the test statistic $Z/\sqrt{V}$ in (3.8), we have

$$Z \approx \frac{1}{2n+1} n \cdot n(\kappa_1 - \kappa_2) = \frac{n^2}{2n+1}(\kappa_1 - \kappa_2),$$

$$V \approx \frac{n \cdot n \cdot 2n}{3(2n+1)^2}[1 - \sum_{k=1}^{K}(\frac{n_{0k} + n_{1k}}{N})^3],$$

where $\kappa_1 = \Phi(\frac{\tau_{s-1} - \mu}{\sigma_0})$ and $\kappa_2 = 1 - \Phi(\frac{\tau_s - \mu}{\sigma_0})$. Since

$$\frac{n \cdot n \cdot 2n}{3(2n+1)^2}[1 - \sum_{k=1}^{K}(\frac{n_{0k} + n_{1k}}{N})^3] \leq \frac{2n^3}{3(2n+1)^2} = V^*,$$

$$\frac{Z}{\sqrt{V}} \geq \frac{Z}{\sqrt{V^*}} = O(\sqrt{n} \cdot (\kappa_1 - \kappa_2)).$$

Then, we have the following lemma.

Lemma 3.1 For the MWM test procedure, the probability of rejecting the null hypothesis converges to 1 if $\kappa_1 \neq \kappa_2$.

Note that the difference, $\kappa_1 - \kappa_2$, plays an important role in this lemma. It can also be regarded as a measure of the skewness of the ordinal variable. The only case where $\kappa_1 - \kappa_2 = 0$ is when $\mu = \frac{\tau_{s-1} + \tau_s}{2}$. That is, when the mean μ falls exactly in the

middle of the category. Otherwise, the probability of rejecting the null hypothesis will be more inflated than the size of the test.

To substantiate the argument that the size of the test is much larger than the nominal value for the MWM test when the treatments have different variances even when the means μ_0 and μ_1 are identical, we perform a small simulation study. Note that the difference in variances implies the violation of the proportional odds assumption. For the simulation study, the ordinal data of the two treatments are generated based on the latent variable model, where the thresholds $(\tau_1, \cdots, \tau_{K-1})$ are fixed at $(-1.5, -0.5, 0.5, 1.5)$. With respect to the latent normal variables, we fix $\sigma_0 = 1$ and $\mu_0 = \mu_1 = \mu$, but several choices of μ and σ_1 are used. The level of significance is chosen to be 0.05. With 10,000 replications, the estimated rate of rejection of the null hypothesis (proportion of rejection of the null hypothesis) is tabulated in Table 3.2. The findings are divided into the following three cases.

(A) Effect of sample size n. We take $\mu = 0.8$ and $\sigma_1 = 3$, representing the case of heterogeneous variances. Hence, the proportional odds model fails to be valid. All reported estimated rates of rejection exceed the level of significance (0.05), and as n increases, the departure from the nominal value increases, reaching 0.6061 for $n = 1000$.

(B) Effect of the difference between σ_1 and σ_0. Several selected values of σ_1 are employed. $\sigma_1 = 1$ is the case of equal proportional odds. As expected, the estimated rate of rejection is close to the level of significance. However, the rate increases rapidly as $\sigma_1 - \sigma_0$ increases.

(C) Effect of the difference between μ and the center of the thresholds. Both σ_1 and n are fixed, which shows that the value of μ also plays a role in the inflation of the estimated rate of rejection, as indicated in the expression of $\kappa_1 - \kappa_2$. As μ moves farther from the center of the thresholds, the effect of μ becomes quite significant in the inflation of the estimated rate of rejection.

Case A		Case B		Case C		Case D	
$\mu = 0.8$, $\sigma_1 = 3$		$\mu = 1$, $n = 500$		$\sigma_1 = 5$, $n = 500$		$\sigma_1 = 1$, $n = 500$	
n	E	σ_1	E	μ	E	μ	E
100	0.1074	1	0.0507	0.0	0.0762	0.0	0.0504
200	0.1832	2	0.2769	0.2	0.1011	0.2	0.0495
400	0.2968	3	0.5483	0.5	0.2217	0.5	0.0474
800	0.5222	5	0.8374	0.8	0.5708	0.8	0.0502
1000	0.6061	7	0.8996	1.0	0.8327	1.0	0.0483

Table 3.2: Estimated rejection rate (E) of the null hypothesis

(D) When the two treatments have both equal means and equal variances, the estimated rates of rejecting the null hypothesis are around the nominal level 0.05.

In conclusion, the simulation study indicates that for non-proportional odds cases, the chance of rejecting the null hypothesis using the MWM testing procedure could be quite large even when the two treatments have the same mean. This indicates the problem of the proportional odds model that have not been addressed in the literature, providing a strong justification for us to search for alternative statistical procedures to compare treatment means when the proportional odds assumption fails to be valid.

3.4 Parameter Estimation of the Latent Variable Model

The latent variable model given in Section 3.2 contains many unknown parameters, and various approaches can be employed to find the parameter estimates. We suggest an estimation method that is easy to implement and is convenient for the further development of multiple comparison procedures.

When the thresholds and all the parameters characterizing the underlying random variables are unknown, the model with the log-likelihood function given by

(3.1) is not identifiable (Poon, 2004). There are different ways to achieve model identification. One can impose constraints on the location and scale of the underlying variable, such as by fixing its mean at 0 and variance at 1, or alternatively one can impose constraints on the thresholds. In the current context, a better strategy is to impose constraints on the thresholds at preassigned values, because the latent means and variances that characterize different treatments are our focus. We now propose the following two-step estimation procedure.

1. Determine the values of the thresholds: let $X_0 \sim N(0,1)$, and use the responses of the control treatment to determine the values pre-assigned to the thresholds.

2. Obtain the estimates of the means and variances that characterize all available treatments: with the thresholds fixed at the values obtained in step 1, estimate the means and variances of the latent variables of all treatments $(\mu_0, \cdots, \mu_G, \sigma_0^2, \cdots, \sigma_G^2)$ under the constraint that the thresholds are the same for all treatments.

Assuming $X_0 \sim N(0,1)$ and using the responses in the control treatment to find the pre-assigned values of the thresholds will produce an estimate of μ_0 that is extremely close to zero and an estimate of σ_0^2 that is extremely close to 1, which serves to fix the locations and scales of the latent variables with reference to those of the control treatment. The values of μ_i and σ_i^2 of other treatments can then be compared to the control in a relative sense. For example, if the estimate of μ_i is greater than 0, the mean difference of the ith treatment and the control is greater than 0.

Note that treatments other than the control can be chosen in step 1 to produce the threshold estimates. For instance, we can let the distribution of X_i be $N(0,1)$ and proceed with the suggested method. Although this will have no effect on the test results, it is not convenient in terms of interpretation. For example, the location parameter of a treatment will then represent the difference between the treatment

and the chosen treatment in step 1, rather than the difference between the treatment and the control.

In multiple comparisons with continuous response, a general assumption is that the samples of different treatments are independent and have a common variance (Dunnett, 1955; Dunnett and Tamhane, 1991, 1992). Our proposed model can easily accommodate the common variance assumption, it is equivalent to assume that $X_0 \sim N(0,1)$, and $X_i \sim N(\mu_i, 1)$, $i = 1, \cdots, G$.

The log-likelihood function for step 1 of the two-step estimation procedure is

$$L_1(\boldsymbol{\tau}) = \sum_{k=1}^{K} n_{0k} \log(\pi_{0k}(\boldsymbol{\tau})) \tag{3.9}$$

where $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_{K-1})'$ is the vector containing all unknown thresholds, and

$$\pi_{0k}(\boldsymbol{\tau}) = \int_{\tau_{k-1}}^{\tau_k} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)dx = \Phi(\tau_k) - \Phi(\tau_{k-1}). \tag{3.10}$$

Based on (3.9) and (3.10), we can derive the MLEs of the thresholds in closed form as follows.

$$\hat{\tau}_k = \Phi^{-1}((n_{01} + \cdots + n_{0k})/n_0), \quad k = 1, \cdots, K-1, \tag{3.11}$$

where $\Phi^{-1}(\cdot)$ is the inverse c.d.f. of the standard normal distribution.

The structure of the log-likelihood function for step 2 is the same as (3.1), but in which the thresholds are fixed values and only $\boldsymbol{\theta}_0$ are unknown parameters. The MLEs of the unknown parameters in step 2 cannot be expressed in a closed form. Therefore, numerical methods have to be applied. One possibility is to use the widely available free software Mx developed by Neale *et al.* (1999) to obtain the estimates in the model.

3.5 Multiple Testing of Several Treatments with Control

3.5.1 Test Statistics

Let $\hat{\eta}_i = \hat{\mu}_i - \hat{\mu}_0$ be the MLE of η_i, $i = 1, \cdots, G$, and $Var(\hat{\eta}_i)$ be the corresponding variance. To derive the test statistics for testing the G null hypotheses stated in (3.3) simultaneously, we need to evaluate the variances and covariances of $\hat{\eta}_i, i = 1, ..., G$.

Theorem 3.2 When the sample size is large, we have

$$Var(\hat{\mu}_i) = \frac{\sigma_i^2}{n_i} \cdot \delta(\mu_i, \sigma_i), \quad i = 0, \cdots, G, \tag{3.12}$$

where

$$\begin{aligned} \delta(\mu_i, \sigma_i) &= \delta_0(\mu_i, \sigma_i) \\ &+ \frac{\delta_0^2(\mu_i, \sigma_i) \cdot \delta_1^2(\mu_i, \sigma_i)}{\delta_2(\mu_i, \sigma_i) - \delta_0(\mu_i, \sigma_i) \cdot \delta_1^2(\mu_i, \sigma_i)}, \end{aligned} \tag{3.13}$$

and

$$\begin{aligned} \frac{1}{\delta_0(\mu_i, \sigma_i)} &= \sum_{k=1}^{K} \frac{1}{\pi_{ik}} \cdot [\phi(\frac{\tau_k - \mu_i}{\sigma_i}) - \phi(\frac{\tau_{k-1} - \mu_i}{\sigma_i})]^2, \\ \delta_1(\mu_i, \sigma_i) &= \sum_{k=1}^{K} \frac{1}{\pi_{ik}} \cdot [\phi(\frac{\tau_k - \mu_i}{\sigma_i}) - \phi(\frac{\tau_{k-1} - \mu_i}{\sigma_i})] \\ &\times [(\tau_k - \mu_i)\phi(\frac{\tau_k - \mu_i}{\sigma_i}) - (\tau_{k-1} - \mu_i)\phi(\frac{\tau_{k-1} - \mu_i}{\sigma_i})], \\ \delta_2(\mu_i, \sigma_i) &= \sum_{k=1}^{K} \frac{1}{\pi_{ik}} \cdot [(\tau_k - \mu_i)\phi(\frac{\tau_k - \mu_i}{\sigma_i}) \\ &\quad -(\tau_{k-1} - \mu_i)\phi(\frac{\tau_{k-1} - \mu_i}{\sigma_i})]^2. \end{aligned}$$

The probability π_{ik} is given by (3.2), and ϕ is the p.d.f. of the standard normal distribution. Note that for the control group with $i = 0$, $\mu_0 = 0$ and $\sigma_0^2 = 1$, we have

$$Var(\hat{\mu}_0) = \frac{1}{n_0} \cdot \delta(0, 1) \tag{3.14}$$

From the proof of Theorem 3.2 (see the Appendix B), for $i \neq j$, $\hat{\mu}_i$ and $\hat{\mu}_j$ are independent. Thus, we have

$$Var(\hat{\eta}_i) = Var(\hat{\mu}_i - \hat{\mu}_0) = \frac{\sigma_i^2}{n_i} \cdot \delta(\mu_i, \sigma_i) + \frac{1}{n_0} \cdot \delta(0,1). \tag{3.15}$$

The following theorem provides the formula to compute the correlations among the test statistics.

Theorem 3.3 When the sample size is large, we have

$$Cov(\hat{\eta}_i, \hat{\eta}_j) = \frac{1}{n_0}\delta(0,1), \quad i \neq j,\ i,j = 1,\ldots,G, \tag{3.16}$$

$$Corr(\hat{\eta}_i, \hat{\eta}_j) = b_i b_j, \tag{3.17}$$

where

$$b_i = \frac{1}{\sqrt{\frac{n_0\sigma_i^2}{n_i} \cdot \frac{\delta(\mu_i,\sigma_i)}{\delta(0,1)} + 1}}.$$

The proof of Theorem 3.3 is also given in the Appendix B. The estimate of the variances and covariances of $\hat{\eta}_i$, $i = 1, ..., G$, can be obtained by replacing $\boldsymbol{\theta}$ in (3.15) and (3.16) by the MLE of $\boldsymbol{\theta}$. Then, the test statistics are

$$Z_i = \frac{\hat{\eta}_i}{\sqrt{\hat{Var}(\hat{\eta}_i)}} = \frac{\hat{\mu}_i - \hat{\mu}_0}{\sqrt{\frac{\hat{\sigma}_i^2}{n_i} \cdot \hat{\delta}(\hat{\mu}_i, \hat{\sigma}_i) + \frac{1}{n_0} \cdot \hat{\delta}(0,1)}}, \tag{3.18}$$

$i = 1, ..., G$. Further, under the null hypotheses these test statistics are distributed approximately as multivariate normal with mean $\mathbf{0}$ and correlation matrix $\{\rho_{ij}\}$, where $\rho_{ij} = 1$ for $i = j$, and $\rho_{ij} = b_i b_j$ for $i \neq j$ as given by (3.17).

3.5.2 Multiple Testing Procedures

Both single-step and stepwise testing procedures can be used to simultaneously test the G null hypotheses (3.3) against the two-sided alternatives (3.4). All testing procedures to be introduced in this chapter satisfy the usual requirements that the familywise type I error rate (FWE), the probability of making at least one type I error, is being controlled at a pre-specified level, say α.

For single-step procedures, each null hypothesis H_i is rejected if the corresponding test statistic Z_i is larger than a single critical value. Two popular single-step testing procedures are examined here.

Bonferroni Procedure (B)

Based on the Bonferroni Inequality, if $|Z_i|$ is larger than $z_{\alpha/2G}$, then the null hypothesis is rejected. The Bonferroni procedure is extremely simple to implement because each test statistic is compared to a single critical value obtained from a standard normal distribution. However, this procedure is quite conservative as G increases. The power of the test drops substantially as compared to the other multiple testing methods to be introduced.

Dunnett Single-step Procedure (DSS)

For multiple comparisons to a control, the widely used Dunnett procedure (1955) is a single-step procedure that compares the test statistics to the critical value $d_{\alpha,G}$, which satisfies the following equation.

$$P(|Z_i| < d_{\alpha,G}, \quad i = 1, ..., G) = 1 - \alpha. \tag{3.19}$$

As the test statistics have a multivariate distribution with a product correlation structure (3.17), the computation of the critical value $d_{\alpha,G}$ is relatively simple. The algorithm is given by Dunnett (1989) and selected values are tabulated by Bechhofer and Dunnett (1988).

Compared to single-step procedures, stepwise procedures provide more powerful testing tools to compare treatments with a control. Here, we discuss two stepwise testing procedures.

Hochberg Procedure (H)

For each hypothesis H_i, let the observed test statistics be z_i, then the p-value of the test for Z_i is

$$p_i = P(|Z_i| > |z_i|). \tag{3.20}$$

For $|z_1|, |z_2|, ..., |z_G|$, let the ordered values be $r_1 \geq r_2 \geq \cdots \geq r_G$. Further, let the ordered p-values be $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(G)}$, and the corresponding hypotheses be $H_{(1)}, H_{(2)}, ..., H_{(G)}$. The Hochberg procedure (1988) is as follows. If FWE is being specified at level α, the testing begins by comparing $p_{(G)}$ to α. If $p_{(G)} \leq \alpha$, then all hypotheses are rejected. If not, then $H_{(G)}$ is accepted and we continue to test $H_{(G-1)}$ by comparing the p-value $p_{(G-1)}$ to $\alpha/2$. If $p_{(G-1)} \leq \alpha/2$, then all hypotheses $H_{(G-1)}, ..., H_1$ are rejected. Otherwise, $H_{(G-1)}$ is retained and we proceed in a similar manner sequentially, comparing $p_{(i)}$ to $\alpha/(G-i+1)$ when $H_{(i)}$ is being tested. The Hochberg procedure is a step-up procedure, testing the least significant hypothesis first and continuing with the more significant hypothesis sequentially. This procedure does not utilize the information of the correlations among the test statistics and is generally less powerful than those that incorporate correlation information in their testing procedures. However, the Hochberg procedure is very simple to implement and more powerful than the single-step procedures (as discussed in the next section).

Dunnett and Tamhane Step-down Procedure (DTSD)

To use the Dunnett and Tamhane (1991) step-down procedure, it is required to compute $d_{\alpha,1}, d_{\alpha,2}, ..., d_{\alpha,G}$. The critical value $d_{\alpha,i}, i = 1, ..., G$, depends on ρ_{ij}. The implementation of the DTSD procedure is much simplified by using the average correlation

$$\bar{\rho} = \frac{2}{G(G-1)} \sum_{1 \leq i < j \leq G} \rho_{ij} \tag{3.21}$$

to obtain the critical constants. This approximation is quite satisfactory unless there is a severe imbalance of sample sizes (Cheung and Chan, 1996).

The DTSD testing procedure is a step-down procedure based on the closure principle of Marcus, Peritz and Gabriel (1976) that begins by comparing r_1 to $d_{\alpha,G}$. If $r_1 \leq d_{\alpha,G}$, then all hypotheses are retained without further testing. If not, then $H_{(1)}$ is rejected and we continue to test $H_{(2)}$ by comparing the r_2 to $d_{\alpha,G-1}$. In general, $H_{(i)}$ is tested by comparing r_i to $d_{\alpha,G-i+1}$. If $r_i \leq d_{\alpha,G-i+1}$, $H_{(i)}, H_{(i+1)}, ..., H_{(G)}$ are

retained and the test stops. Otherwise, $H_{(i)}$ is rejected and the test continues with $H_{(i+1)}$.

In the literature, there are variations similar to the aforementioned testing procedures. For example, Rom (1990) modified the Hochberg procedure and suggested a step-up procedure that is slightly more powerful. However, the improvement is far from substantial. Another example is the Dunnett and Tamhane (1992) step-up procedure. Even though it yields slightly higher power than the DTSD procedure, the computation of the necessary critical values is quite complex (Kwong and Liu, 2000). Hence, for practical purposes, we believe that the procedures introduced here are adequate. Next, the power of the four testing procedures introduced in this section are compared.

3.5.3 Power Comparison: a Simulation Study

There are several possible definitions of power in the literature on multiple testing. Horn and Dunnett (2004) discussed several different definitions, such as, all-pairs power, any-pair power, per-pair power, and average power (the proportion of false hypotheses that are correctly rejected). In this section, we report a simulation study that compares the four testing procedures (B, DSS, H, and DTSD) in terms of average power.

For our simulation study, the ordinal data of different treatments are generated based on the latent variable model with thresholds fixed at $(-1.5, -0.5, 0.5, 1.5)$, and the latent variables are assumed to be $X_0 \sim N(0,1)$ and $X_i \sim N(\mu_i, 1)$, $i = 1, \cdots, G$. The estimation of parameters follows the two-step estimation procedure given in section 3.4, and the required information for evaluating the variances and covariances of $\hat{\eta}_i$ is given in Theorem 3.2. In the simulation, we focus on the behavior of the average power of each testing procedure for different sample size configurations and different true/false configurations.

1. The average power for different sample size configurations

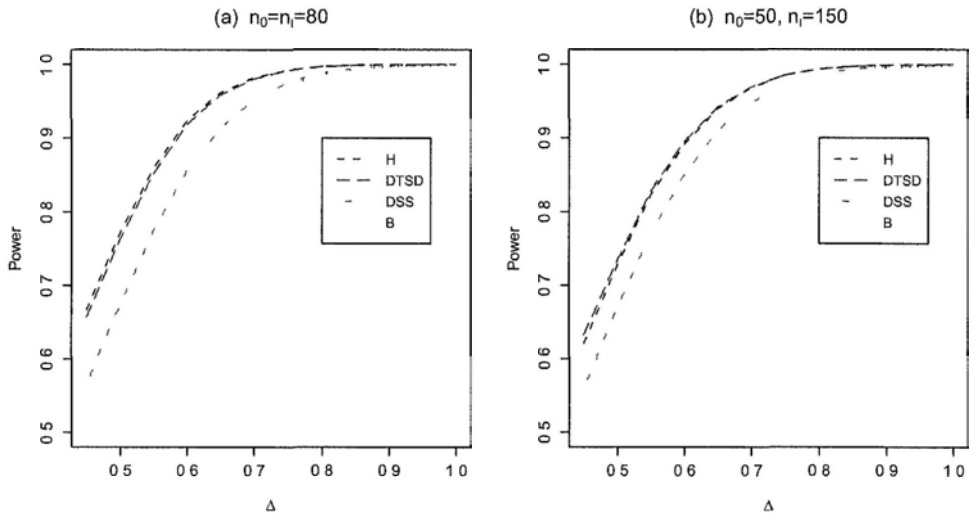

Figure 3.1: Average power of H, DTSD, DSS, and B.

For illustrative purposes, G is chosen to be 5 and let $\eta_i = \mu_i - \mu_0 = \Delta$ for all $i = 1, ..., 5$. The simulated average power is computed for a wide range of Δ. Two different patterns of sample size configurations are selected: (a) $n_0 = n_1 = \cdots = n_5 = 80$ and (b) $n_0 = 50, n_1 = \cdots = n_5 = 150$. Note that the correlations among the test statistics are higher for the second pattern from (3.17), and we expect that the DTSD procedure performs better than the H procedure because it utilizes the correlation structure information in its testing algorithm. The estimated average power is evaluated based on 100,000 replications. Figure 3.1 presents the average power of the four different procedures for the selected sample size configurations. To produce an informative comparison (Figure 3.2), using the B procedure as the baseline, the percentage increase in the simulated average power is computed for the H, DTSD, and DSS procedures.

The important findings are summarized as follows.

(a) The procedures H, DTSD, and DSS have uniformly higher average power than the B procedure because the B procedure is quite conservative.

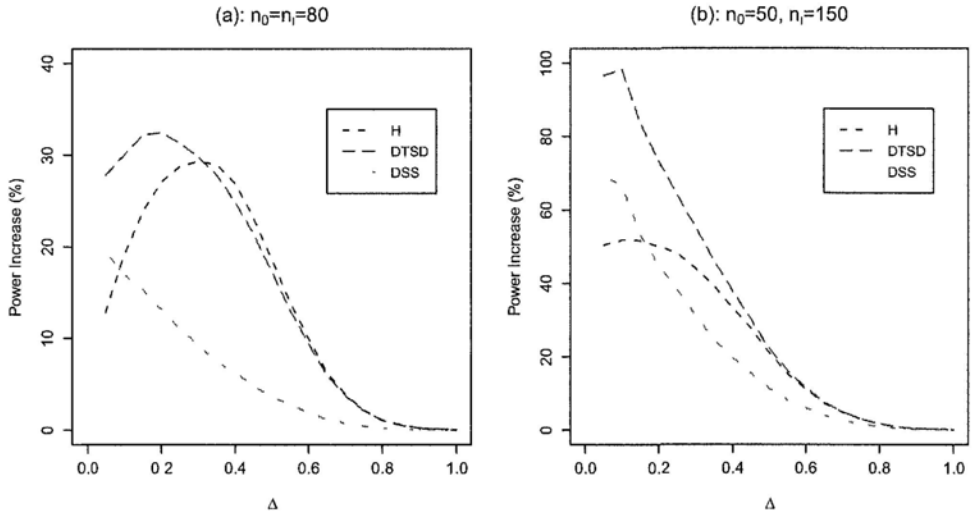

Figure 3.2: Increase in average power of H, DTSD, and DSS as compared to B.

(b) Stepwise procedures are generally more powerful than single-step procedures. In fact, it is easy to see that the DTSD dominates the DSS procedure due to the critical values being used ($d_{\alpha,G} \geq d_{\alpha,i},\ i = 1, ..., G$). The H procedure is also more powerful than the DSS procedure except when Δ is very small (less than 0.1 for equal sample sizes, and less than 0.2 for the case in which the sample sizes are unequal).

(c) Correlations among the test statistics play an important role in the differences of power among the various procedures. Let us compare the increase in power given in Figure 3.2 for the two different sample size configurations. For the equal sample size configuration, the average correlation of the test statistics is about 0.5, whereas it is about 0.75 for the unequal sample size case. It is worth noting when compare Figure 3.2 (a) with Figure 3.2 (b), the relative power of the H procedure as compared to the DSS and the DTSD procedures decreases substantially because the H procedure does not utilize the correlation information of the test statistics.

(d) In summary, we propose to use the DTSD procedure which performs relatively better than the other procedures unless the correlations among the test statistics are only moderate. In those cases, the H procedure can serve as an alternative due to its simplicity.

Similar findings can be obtained using the other concepts of power and are thus not reported here. When $\Delta = 0$, all the procedures introduced here were found to control the FWE at level α, including the cases when $\sigma_i \neq \sigma_0$, $i = 1, \cdots, G$, according to our simulation study. This is an expected result and hence the findings are not tabulated.

2. The average power for different true/false configurations

It will be informative to study the influence of the true/false configurations to the power, since the number of false hypotheses are usually unknown in the practical applications; especially in the case to determine the sample size that guarantee a specified power, people have to look for the least favorable configuration (LFC) where the power attains its minimum.

For illustration, we still take $G = 5$, and let $\Delta = 0.2$ for false hypotheses. We assume all treatments have equal sample size and $n_0 = n_i = 500$. Based on 100,000 replications, Figure 3.3 presents the average power of the four testing procedures for different true/false configurations.

From Figure 3.3, the single-step procedures (B and DSS) will have equal average power for different true/false configurations; the stepwise procedures (H and DTSD) will attain their minimum average power when only one false hypothesis, and the average power will increase when the number of false hypotheses are increasing. This behavior of the average power for different true/false hypotheses is quite different from that of the all-pairs power, which is defined as the probability of rejecting all false hypotheses. Figure 3.4 gives the plot of the all-pairs powers of the four testing procedures under the same simulation setting as in Figure 3.3. It can be found that the stepwise procedures will attain their minimum all-pairs power in some intermediate value of F (the number of false hypotheses), and the single-step

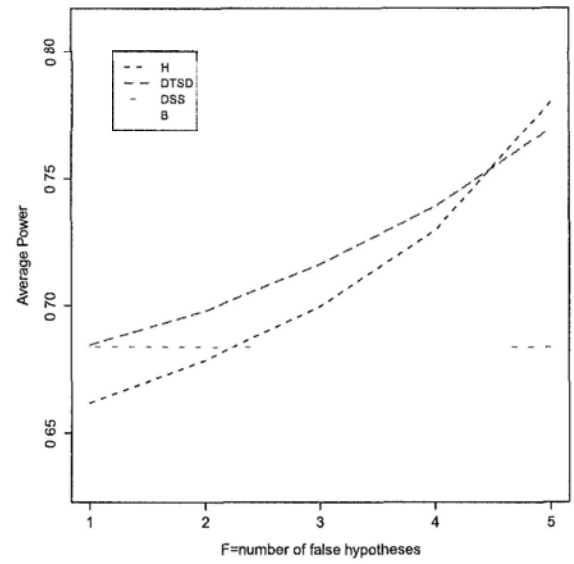

Figure 3.3: Average power of H, DTSD, DSS, and B for different true/false configurations.

procedures will have downward trend as the increase of F.

3.6 Examples

3.6.1 Example 1

This example is extracted from Fujii and Itakura (2009). A randomized, double-blind, placebo-controlled study was conducted to compare the efficacy of intravenous pretreatment with fentanyl 50 μg, fentanyl 100 μg, and lidocaine 40 mg, preceded by venous occlusion, for reducing pain on the injection of propofol in Japanese surgical patients. Patients were interviewed to assess pain intensity on injection using a 4-point verbal rating scale (0=none, 1=mild, 2=moderate, 3=severe). The ordinal categorical observations of the four treatments are given in Table 3.3.

Based on our two-step estimation procedure, the thresholds are fixed at $(\hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3)$

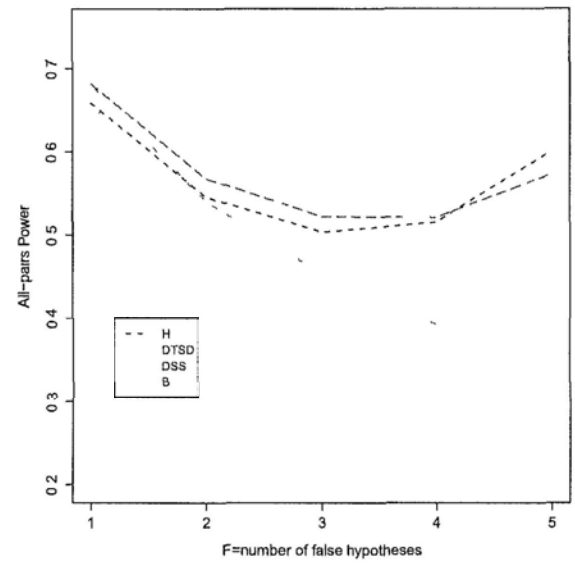

Figure 3.4: All-pairs power of H, DTSD, DSS, and B for different true/false configurations.

= (-0.9674, 0.0837, 0.9674), the estimated means $(\hat{\mu}_0, \cdots, \hat{\mu}_3)$ = (0.0, -0.2617, -1.5478, -1.6096), and the estimated standard deviations $(\hat{\sigma}_0, \cdots, \hat{\sigma}_3)$ = (1.0, 1.0313, 1.7616, 1.5482). To simultaneously test the difference between each treatment and the control, the hypotheses are

$$H_i : \mu_i - \mu_0 = 0 \quad \text{vs.} \quad H_i' : \mu_i - \mu_0 \neq 0, \qquad i = 1, 2, 3,$$

with familywise error rate $\alpha = 0.05$. Using the procedures given in section 5, the test statistics (z_1, z_2, z_3) = (-0.9320, -2.9070, -3.2059), and the corresponding two-sided p-values (p_1, p_2, p_3) = (0.3514, 0.0036, 0.0013). The average correlation of the test statistics for this example is $\bar{\rho} = 0.2209$. The critical values for the ordered test statistics in the DTSD procedure are 1.9600, 2.2321, and 2.3808. All of the testing procedures given in the previous section produce the same conclusion. That is, fentanyl 100 μg and Lidocaine 40 mg are able to reduce the pain, while fentanyl 50 μg is ineffective in reducing the pain compared with a placebo.

	Grading of pain				Sample
Treatment	0	1	2	3	size
Placebo	5	11	9	5	30
Fentanyl 50 μg	7	13	6	4	30
Fentanyl 100 μg	19	5	4	2	30
Lidocaine 40 mg	20	5	4	1	30

Table 3.3: Incidence and intensity of pain on injection of propofol

For this example, we can consider the estimates of the log-odds ratios based on (3.5) when we compare each treatment with the control separately. For example, when we compare fentanyl 100 μg with the placebo based on the observations in Table 3.3, the estimated log-odds ratios $\hat{\gamma}_k$, $k = 1, 2, 3$, are 2.16, 1.25, and 1.03, respectively, indicating that the adoption of the proportional odds assumption may be questionable.

3.6.2 Example 2

This example is taken from Koo *et al.* (2006). In a prospective, randomized, double-blind, placebo-controlled study to reduce the pain of propofol injection, 240 patients representing for elective surgery were randomly allocated into eight groups. The pain scores of the patients were assessed using the verbal rating scale (0=no pain, 1=mild pain or soreness, 2=moderate pain, and 3=severe pain). Various treatments were compared to the control (saline). The observations are presented in Table 3.4.

Based on our two-step estimation procedure, the thresholds are fixed at $(\hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3)$ = (-1.1108, 0.1679, 1.5011). The estimated means $(\hat{\mu}_0, \cdots, \hat{\mu}_7)$ = (0.0, -1.0290, -1.1878, -0.8117, -0.7029, -1.6391, -0.2154, -0.4331), and the estimated standard deviations $(\hat{\sigma}_0, \cdots, \hat{\sigma}_7)$ = (1.0, 0.9101, 1.0187, 0.8547, 1.1993, 2.0072, 0.9296, 1.6641).

We simultaneously test the following hypotheses

$$H_i : \mu_i - \mu_0 = 0 \quad \text{vs.} \quad H_i' : \mu_i - \mu_0 \neq 0, \quad i = 1, \cdots, 7,$$

with FWE = 0.05. Based on the estimated values of the parameters, the test

Intensity of pain	Treatment							
	S	L	K100	K50	K10	M	KP	Pre
None (=0)	4	14	16	11	11	18	4	10
Mild (=1)	13	13	11	15	12	7	18	10
Moderate (=2)	11	3	3	4	6	3	6	6
Severe (=3)	2	0	0	0	1	2	2	4

S=saline, L=lidocaine, K100=ketamine 100 $\mu g/kg$, K50=ketamine 50 $\mu g/kg$, K10=ketamine 10 $\mu g/kg$, M=midazolam premedication, KP=ketamine 100 $\mu g/kg$ in propofol solution following saline 2 ml, Pre=pretreatment with ketamine 100 $\mu g/kg$ 3 min before the injection of propofol.

Table 3.4: Incidence and intensity of pain on injection of propofol

statistics $(z_1, \cdots, z_7)$ = (-3.6558, -3.8281, -3.0520, -2.2262, -2.8878, -0.7982, -1.1145), and the corresponding two-sided p-values $(p_1, \cdots, p_7)$ = (0.0003, 0.0001, 0.0023, 0.0260, 0.0039, 0.4248, 0.2651). To conduct the multiple testing procedures, the ordered p-values or test statistics should be compared with the corresponding critical values. To derive the critical values for the test statistics, the average correlation can be calculated based on the estimation results and equations (3.17) and (3.21). For this example, the average correlation is $\bar{\rho} = 0.3706$. The critical values for the ordered test statistics in the DTSD procedure are 1.9600, 2.2237, 2.3675, 2.4653, 2.5390, 2.5978, and 2.6465.

For this example, all four multiple testing procedures give the following consistent testing conclusions. Compared to the control in terms of pain reduction, only treatments L, K100, K50, and M are significantly different. Among the effective treatments, K100 and L are the most effective treatments and have similar efficacy from an examination of their corresponding p-values.

Even though the testing conclusions are the same with these four procedures, it is easy to recognize that on average that is not the case. For example the critical constants for the stepwise comparisons for the DTSD procedure are 1.9600, 2.2237, 2.3675, 2.4653, 2.5390, 2.5978, and 2.6465, while the critical constant being used to test every hypothesis in DSS is 2.6465. Therefore, on average, it is easy to see that

the DTSD procedure rejects more hypotheses than the DSS procedure.

3.7 Conclusion

In this chapter, we consider the problem of multiple comparison with a control with ordinal responses. The WMW test that relies on the proportional odds model to analyze ordered categorical responses may produce undesirable results when the proportional odds assumption fails to be valid. As indicated by our simulation and theoretical justification, the violation of the proportional odds assumption will lead to top high a probability of claiming a difference in treatment efficacy when in fact there is none.

To rectify this problem, an alternative method is proposed. By considering the ordinal responses as manifestations of underlying continuous variables, we suggest the latent variable model, which facilitates comparisons of the mean efficacy of treatments and accommodates the possibility of heterogeneous variances among different treatments. Both single-step and stepwise multiple testing procedures are examined, and the major contributions of the proposed methods are as follows.

(a) Even though the proportional odds model has been widely applied, as many authors have indicated, the proportional odds assumption may not be valid. The undesirable consequence of the invalid proportional odds assumption is demonstrated in Section 3.3. The inflated probability of rejecting the true null hypothesis about the equivalence of treatment efficacy provides major support for our motivation to seek alternative methods of analyzing ordinal categorical data in such circumstances.

(b) The latent variable model approach is proposed in this chapter to compare several treatments with a control. The two-step estimation procedure is given for the identification and estimation of the latent variable model.

(c) To conduct multiple testing, the test statistics are derived. In particular,

we derive the formulae for the correlations among the test statistics that are necessary for the testing procedures.

(d) Based on the developed test statistics, several multiple testing procedures are introduced for the comparison of several treatments with a control. The merits of these procedures are compared and discussed and we recommend the use of the DTSD procedure which performs better than the others, especially when the correlations among the statistics are large.

We proposed the use of a two-step procedure to estimate the model parameters. The procedure is easy to implement and is accessible to users. Besides the two-step procedure, we have also explored the performance of another iterative approach (see Algorithm 5.1 in Chapter 5). Our findings indicate that this iterative method converges extremely fast and the final estimates are extremely close to the estimates obtained with the two-step procedure. Hence, for computational simplicity, the two-step procedure is sufficient to provide satisfactory estimates.

Chapter 4

A Unified Framework for Treatment Comparisons

4.1 Introduction

In this chapter, we continue to study the comparison of treatments with ordered categorical responses. We established a unified framework that allows various procedures be recognized from a common perspective.

The comparison of two independent treatments (or a treatment and a control) should be the most fundamental task in treatment comparison study. The Wilcoxon-Mann-Whitney (WMW) test (Wilcoxon 1945, Mann and Whitney 1947) may be the most popular nonparametric method used to investigate the effect between two treatments that have ordered categorical responses. As a distribution-free method, the WMW test is also widely used to compare treatments with continuous responses. A comprehensive study of the WMW test is given by Lehmann (1975).

When the responses have continuous distributions, it has been well recognized that the WMW test can provide an exact test of location when the two populations are identical in scale (see e.g. Wetherill, 1960). Wetherill (1960) pointed out that the difference in dispersion and the skewness of the distributions may have significant effect on the behaviour of the WMW test. The research of Pratt (1964) also showed that the level of the WMW test can not be preserved when the populations differ in

dispersion. However, when the responses are ordinal, there is difficulty in directly interpreting the location and dispersion of the responses.

For ordered categorical response, it is in many cases reasonable to consider the ordinal response as the manifestation of an underlying continuous variable (see e.g. McCullagh, 1980; Anderson, 1984). The most popular assumption for the distributions of the underlying continuous random variables is either normal distribution or logistic distribution. On the basis of such a latent variable model, it is convenient to interpret the location and dispersion of the ordinal responses, and many statistical inferences can also be conveniently conducted. The proportional odds model proposed by McCullagh (1980) has been widely adopted in the literature. To compare treatments with ordered categorical responses, Whitehead (1993) proposed to use the log-odds ratio as a measure of mean treatment effect under the proportional odds assumption, and used the WMW test to investigate the treatment effect. By considering the ordinal response as the manifestation of underlying normally distributed random variable, Poon (2004) proposed a method to examine the possible treatment effect, which can be conveniently attributed to either location effect or dispersion effect based on the latent variable model.

As we have studied in Section 3.3, the level of the WMW test with the alternative specified as proportional odds may approach to 1 when the underlying two distributions differ significantly in dispersion, even when they have the same location. This finding, similar to Wetherill (1960) and Pratt (1964), indicates that the WMW test for ordered categorical responses can provide effective test of location only when the two treatments have identical scale in terms of the latent variable model.

In the planning stage of an experiment, it is a very important issue to determine the required sample size to detect a significant treatment effect. In the literature of comparing two treatments with ordered categorical responses, several sample size determination methods have been proposed. For example, Whitehead (1993) gave a sample size formula that is derived based on the WMW test with the alternative

specified as proportional odds. Zhao, Rahardja, and Qu (2008) gave the sample size calculation for the WMW test with the alternative specified as the probability of one treatment being superior to the other. Although these methods used different measures to quantify the treatment effect, they can be interpreted in a unified framework (see Section 4.2).

It should also be widely recognized that the sample size determination methods are closely related to the corresponding testing methods. In other words, the assumptions for the testing methods should also be satisfied when corresponding sample size determination methods are utilized. As the WMW test can provide exact test of location difference only when the two treatments have identical scales, the sample size determination methods (Whitehead, 1993; Zhao, Rahardja, and Qu, 2008) that are based on the WMW test might be questionable when the two treatments have difference scales (see the study in Section 4.6). This motivates us to find new testing and sample size determination method that can accommodate the difference in the scales of the two treatments.

In this chapter, we propose a general analysis framework for the latent variable model, which can be conveniently utilized to compare treatments with ordinal responses. The underlying continuous random variables are allowed to have distributions in a large family, the location-scale distribution family. This family contains some very important distributions, such as normal distribution, logistic distribution, and Cauchy distribution. Thus, our latent variable model will cover the mostly adopted latent variable models in the literature, the latent normal distribution model and the latent logistic distribution model, which usually have good interpretations and applications.

Based on such latent variable model, different treatment effect measures for ordinal responses can be considered in a unified manner. A two-step procedure is proposed for the identification and estimation of the latent variable model, where the location and scale parameters that characterize different treatments can be freely

estimated. Based on the proposed latent variable method, further statistical inferences are also provided. Subsequently, new sample size determination method that can accommodate the scale difference in different treatments is proposed. The newly proposed method is compared with the existing methods in the aspects of power and sample size determination. The problem of the existing sample size determination methods is also investigated in real example and numerical study.

4.2 A Unified Consideration of the Treatment Effect Measures

Suppose we observe two independent treatments with ordinal responses. The ordinal responses of the treatments are classified into one of the K ordered categories C_k, $k = 1, \cdots, K$. Let π_{ik} be the probability that the ordinal response of treatment i falls into category C_k, $i = 1, 2$, $k = 1, \cdots, K$. Let $\gamma_{ik} = \pi_{i1} + \cdots + \pi_{ik}$ be the cumulative probability of treatment i up to category k, $k = 1, \cdots, K-1$.

Following the arguments of McCullagh (1980) and Bartholomew (1980), we may consider the ordered categorical responses as manifestations of some underlying continuous random variables. In this chapter, we assume the underlying continuous random variables have cumulative distribution function $F(x; \mu, \sigma)$, which belongs to the location-scale (L-S) family. If F is a member of the L-S family, then $G(x) = F(\mu + \sigma x)$ is also a cumulative distribution function of the member of the L-S family. So, we may consider the cumulative distribution function $F(x; 0, 1)$ with location 0 and scale 1 as the standard distribution in the L-S family, and simply denote it by $F_0(x)$. The probability density function of $F_0(x)$ will be denoted by $f_0(x)$. The L-S family includes many important distributions, such as the normal distribution, the logistic distribution, and the Cauchy distribution. These distributions usually have good properties and wide applications. For example, if $F(x; \mu, \sigma)$ is the cumulative

logistic distribution function with location μ and scale σ, then

$$F(x;\mu,\sigma) = \frac{1}{1+e^{-(x-\mu)/\sigma}}, \quad \text{and } F_0(x) = \frac{1}{1+e^{-x}}.$$

The inverse function of $F_0(x)$ is the usual logit function. Specifically, for $\gamma \in (0,1)$,

$$F_0^{-1}(\gamma) = \text{logit}(\gamma) = \log(\frac{\gamma}{1-\gamma}).$$

Treatment i is related to the distribution function $F(x;\mu_i,\sigma_i)$ by a set of thresholds $\tau_1 < \cdots < \tau_{K-1}$, which have the same (but unknown) values for both treatments, such that

$$\gamma_{ik} = F_0(\frac{\tau_k - \mu_i}{\sigma_i}), \quad i = 1,2,\ k = 1,\cdots,K-1, \tag{4.1}$$

where $F_0(x) = F(x;0,1)$ is the standard distribution function of the L-S family. Equivalently, (4.1) can be written as

$$F_0^{-1}(\gamma_{ik}) = (\tau_k - \mu_i)/\sigma_i.$$

Here, $F_0^{-1}(\cdot)$ can be regard as a link function, which can be expressed in the following general form

$$\text{link}(\gamma_{ik}) = (\tau_k - \mu_i)/\sigma_i, \quad i = 1,2,\ k = 1,\cdots,K-1. \tag{4.2}$$

If the covariates of treatment i are considered, we only need to let $\mu_i = \beta^T \mathbf{x}_i$, where $\mathbf{x}_i$ is the covariate vector and β is the coefficients.

A special case is that the two underlying continuous distributions have equal scale parameters. Suppose $\sigma_1 = \sigma_2 = \sigma$, then in this case,

$$\text{link}(\gamma_{ik}) = F_0^{-1}(\gamma_{ik}) = (\tau_k - \mu_i)/\sigma. \tag{4.3}$$

Then the treatment effect between treatment 1 and treatment 2 can be expressed as

$$\Delta_k = \text{link}(\gamma_{2k}) - \text{link}(\gamma_{1k}) = (\mu_1 - \mu_2)/\sigma = \Delta, \quad k = 1,\cdots,K-1. \tag{4.4}$$

This unified L-S framework is general enough to align many popular models for analyzing ordinal categorical variables. The proportional odds model of McCullagh (1980) is a special case of (4.3) with the "link" taken as the logit function (or equivalently, F_0 taken as the standard logistic distribution function). The treatment effect in terms of log-odds ratio, denoted by Δ_{log}, is given by

$$\Delta_{log} = \text{logit}(\gamma_{2k}) - \text{logit}(\gamma_{1k}) = \log\left(\frac{\gamma_{2k}(1-\gamma_{1k})}{\gamma_{1k}(1-\gamma_{2k})}\right) = (\mu_1 - \mu_2)/\sigma. \tag{4.5}$$

This is also the model with common log-odds ratio assumption used by Whitehead (1993), where the common log-odds ratio, denoted by Δ_{whd}, is defined as

$$\Delta_{whd} = \log\left(\frac{\gamma_{2k}(1-\gamma_{1k})}{\gamma_{1k}(1-\gamma_{2k})}\right) \quad \text{for all } k.$$

The afore discussion allows us to have a new interpretation of the proportional odds model. The proportional odds model is a special case of the latent variable model in (4.1), and is obtained when (a) the distributions are logistic; (b) the logistic distributions for all treatments have equal scales. This interpretation for the proportional odds assumption has also been mentioned by Peterson and Harrell (1990). The test of the proportional odds assumption can be conducted based on the two interpretations. In fact, they can be conveniently performed by the tests (I) and (II) respectively in Section 4.3.2. The test of interpretation (b) is crucial, because the violation of this assumption will lead to questionable inference results (see the study in Section 4.6). The existing test methods for the proportional odds assumption did not depend on such an interpretation (see e.g. Brant, 1990; SAS Proc Logistic). Thus, our testing method based on the new interpretation of proportional odds model provides a useful supplement to the significance tests for the proportional odds assumption.

When F_0 in (4.1) is taken as the standard normal distribution function Φ (or equivalently, taken the "link" in (4.2) as probit function), we obtain the latent normal distribution model of Poon (2004). Note that the model based on (4.1) or (4.2) is not identifiable. Some constraints must be imposed to the parameters to

achieve model identification. Poon (2004) gave a maximum likelihood estimation of the unknown parameters with some constraints imposed. Based on Poon's (2004) model, the treatment effect can be attributed to either location effect or dispersion effect. When the underlying normal distributions have equal variance, say σ, the treatment effect, denoted by Δ_{norm}, can be directly given by

$$\Delta_{norm} = \Phi^{-1}(\gamma_{1k}) - \Phi^{-1}(\gamma_{2k}) = (\mu_2 - \mu_1)/\sigma. \tag{4.6}$$

The Wilcoxon-Mann-Whitney test is also a widely used method to compare two treatments with ordered categorical responses. This WMW test interprets the treatment effect as the probability of one treatment being superior to another treatment. Specifically, let Y_1 be the ordinal variable corresponding to treatment 1, and Y_2 be the ordinal variable corresponding to treatment 2. Then, the treatment effect measure, denoted by Δ_{wmw}, is given by

$$\Delta_{wmw} = P(Y_1 < Y_2) + 0.5P(Y_1 = Y_2) \tag{4.7}$$

The equivalence of the two treatments corresponds to $\Delta_{wmw} = 0.5$. Let $\boldsymbol{\gamma}_1 = (\gamma_{11}, \cdots, \gamma_{1(K-1)})'$ and $\boldsymbol{\gamma}_2 = (\gamma_{21}, \cdots, \gamma_{2(K-1)})'$. The treatment effect measure Δ_{wmw} can also be expressed as (Ryu and Agresti, 2008)

$$\Delta_{wmw} = \boldsymbol{\gamma}_1' D \boldsymbol{\gamma}_2 + 0.5(1 + \gamma_{1(K-1)} - \gamma_{2(K-1)}), \tag{4.8}$$

where

$$D = \begin{pmatrix} 0 & 0.5 & 0 & 0 & \cdots & 0 \\ -0.5 & 0 & 0.5 & 0 & \cdots & 0 \\ & \vdots & & & \vdots & \\ 0 & \cdots & 0 & -0.5 & 0 & 0.5 \\ 0 & \cdots & 0 & 0 & -0.5 & 0 \end{pmatrix}.$$

Thus, the treatment effect measure Δ_{wmw} is expressed as a function of γ_{ik}, and the γ_{ik} can be linked with the underlying distribution in the form of (4.3). When the parameter form of the underlying distribution is specified, the estimate of Δ_{wmw}

can be derived based on the maximum likelihood method. Note that the WMW test does not depend on a specific form of the underlying distribution, but it can provide exact location test only when the underlying distributions have equal scale parameters.

4.3 A Framework for the Analysis of the Latent Variable Model

4.3.1 Model Estimation

The latent variable model is widely used for the analysis of ordered categorical responses, see e.g. Anderson and Philips (1981), Qu, Piedmonte and Medendorp (1995), Bekele and Thall (2004), Todem, Kim, and Lesaffre (2007), Leon-Novelo *et al.* (2010). A common feature of these models in the literature is that the scale parameters of the underlying distributions are either assumed to be identical or fixed at some specific values. We now outline our proposed latent variable model as follows, which is more flexible and allows for heterogeneous scales across different treatments.

Following the idea of the above section, we still relate the observed ordinal response with a underlying continuous random variable with distribution belonging to the L-S family. We generalize the two-treatment model and consider G independent treatments with ordinal responses. We continue to assume that treatment i, $i = 1, \cdots, G$, is related to the underlying random variable X_i with distribution $F(x; \mu_i, \sigma_i)$ through a set of thresholds $\tau_1 < \cdots < \tau_{K-1}$ as in (4.1). Or equivalently, the ordinal response of treatment i falls into category C_k, $k = 1, \cdots, K$, if and only if $\tau_{k-1} < X_i < \tau_k$, where $\tau_0 = -\infty$ and $\tau_K = \infty$. We again assume a common set of thresholds for all treatments.

Let n_i be the total number of subjects receiving treatment i, and n_{ik} be the number of subjects receiving treatment i with outcomes classified into category C_k.

In addition, let $N = \sum_{i=1}^{G} n_i$ be the total sample size. Then, the log-likelihood function for the samples of the G treatments is given by

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{G} \sum_{k=1}^{K} n_{ik} \log(\pi_{ik}(\boldsymbol{\theta})), \tag{4.9}$$

where

$$\pi_{ik}(\boldsymbol{\theta}) = F_0(\frac{\tau_k - \mu_i}{\sigma_i}) - F_0(\frac{\tau_{k-1} - \mu_i}{\sigma_i}).$$

Here, $\boldsymbol{\theta} = (\boldsymbol{\tau}', \boldsymbol{\theta}_0')'$ is a vector containing all the unknown parameters, with $\boldsymbol{\tau}' = (\tau_1, \cdots, \tau_{K-1})$ and $\boldsymbol{\theta}_0' = (\mu_1, \cdots, \mu_G, \sigma_1, \cdots, \sigma_G)$.

Under the specification of the latent variable model, the dispersion effect of the G treatments can be investigated by testing the following null hypothesis

$$H_0 : \sigma_1 = \cdots = \sigma_G. \tag{4.10}$$

Similarly, the location effect of the G treatments can be investigated by testing the null hypothesis

$$H_0 : \mu_1 = \cdots = \mu_G. \tag{4.11}$$

Here, the parameters μ_i and σ_i that characterize different treatments are our interest of study. However if all the parameters involved in (4.9) are unknown, the latent variable model is not identifiable. Following the arguments as in Section 3.4, we propose the following two-step procedure to identify the model and produce parameter estimates.

Step 1. Determine the values of the thresholds: let $X_s \sim F(x; 0, 1)$, $1 \leq s \leq G$, and use the responses of treatment s to determine the values of the thresholds.

Step 2. Obtain the estimates of the location and scale parameters that characterize different treatments: with the thresholds fixed at the values obtained in step 1, estimate μ_i and σ_i, $i = 1, \cdots, G$.

This two-step estimation procedure with the thresholds determined by treatment s is equivalent to specify treatment s as a reference with location 0 and scale 1 for

treatment comparison. Note that such specification of treatment s as a reference is achieved not directly through fixing the parameters of treatment s but by fixing the thresholds. This will avoid the estimation of the complicated correlation structure between the thresholds and the other parameters, and thus make following inferences more convenient. The treatment comparison as stated in (4.10) and (4.11) can then be performed in a relative sense. We have derived a theoretical result that the selection of different treatment to determine the thresholds in step 1 has no effect on the testing results for large sample size (see Lemma 4.1). A simulation study is also conducted to study the performance when the sample size is small. It is found that the testing results are almost the same when use different treatments to determine the thresholds in step 1, and the difference is negligible. Without loss of generality, in this chapter, we use treatment 1 (s=1) to determine the thresholds.

The log-likelihood function for the step 1 estimation procedure is

$$L_1(\boldsymbol{\tau}) = \sum_{k=1}^{K} n_{sk} \log(\pi_{sk}(\boldsymbol{\tau})), \tag{4.12}$$

where

$$\pi_{sk}(\boldsymbol{\tau}) = F_0(\tau_k) - F_0(\tau_{k-1}).$$

The MLE of $\boldsymbol{\tau}$ can be given in closed form by

$$\hat{\tau}_k = F_0^{-1}(\frac{n_{s1} + \cdots + n_{sk}}{n_s}), \quad k = 1, \cdots, K-1. \tag{4.13}$$

The log-likelihood function for the step 2 estimation procedure can be written as

$$L_2(\boldsymbol{\theta}_0) = \sum_{i=1}^{G} \sum_{k=1}^{K} n_{ik} \log(\pi_{ik}(\boldsymbol{\theta}_0)), \tag{4.14}$$

where

$$\pi_{ik}(\boldsymbol{\theta}_0) = F_0(\frac{\tau_k - \mu_i}{\sigma_i}) - F_0(\frac{\tau_{k-1} - \mu_i}{\sigma_i}). \tag{4.15}$$

The MLE of $\boldsymbol{\theta}_0$, denoted by $\hat{\boldsymbol{\theta}}_0$, involved in (4.14) can not be derived in a closed form. Numerical methods must be applied. The score function and the Fisher information matrix of the log-likelihood function (4.14) are derived in the Appendix D.

So, the Newton-Raphson method can be applied efficiently to obtain $\hat{\boldsymbol{\theta}}_0$. From the Appendix D, we can also derive the variances of the estimates in a closed form. Specifically,

$$Var(\hat{\mu}_i) = \frac{\sigma_i^2}{n_i} \cdot \delta(\mu_i, \sigma_i), \quad i = 1, \cdots, G, \tag{4.16}$$

where,

$$\delta(\mu_i, \sigma_i) = \delta_0(\mu_i, \sigma_i) + \frac{\delta_0^2(\mu_i, \sigma_i) \cdot \delta_1^2(\mu_i, \sigma_i)}{\delta_2(\mu_i, \sigma_i) - \delta_0(\mu_i, \sigma_i) \cdot \delta_1^2(\mu_i, \sigma_i)}, \tag{4.17}$$

and

$$\begin{aligned}
\frac{1}{\delta_0(\mu_i, \sigma_i)} &= \sum_{k=1}^{K} \frac{1}{\pi_{ik}} \cdot [f_0(\frac{\tau_k - \mu_i}{\sigma_i}) - f_0(\frac{\tau_{k-1} - \mu_i}{\sigma_i})]^2, \\
\delta_1(\mu_i, \sigma_i) &= \sum_{k=1}^{K} \frac{1}{\pi_{ik}} \cdot [f_0(\frac{\tau_k - \mu_i}{\sigma_i}) - f_0(\frac{\tau_{k-1} - \mu_i}{\sigma_i})] \\
&\qquad \times [(\tau_k - \mu_i) f_0(\frac{\tau_k - \mu_i}{\sigma_i}) - (\tau_{k-1} - \mu_i) f_0(\frac{\tau_{k-1} - \mu_i}{\sigma_i})], \\
\delta_2(\mu_i, \sigma_i) &= \sum_{k=1}^{K} \frac{1}{\pi_{ik}} \cdot [(\tau_k - \mu_i) f_0(\frac{\tau_k - \mu_i}{\sigma_i}) - (\tau_{k-1} - \mu_i) f_0(\frac{\tau_{k-1} - \mu_i}{\sigma_i})]^2.
\end{aligned}$$

In these expressions, π_{ik} is given by (4.15), and $f_0(x)$ is the p.d.f. of standard distribution $F_0(x)$.

Note that, under the two-step estimation procedure, for the treatment s with $\mu_s = 0$ and $\sigma_s = 1$, we have

$$Var(\hat{\mu}_s) = \frac{1}{n_s} \cdot \delta(0, 1). \tag{4.18}$$

In the context of comparing only two treatments, the MLE of Δ_{norm} in (4.6) can be obtained via the aforementioned framework, with the F_0 in (4.15) chosen as the standard normal distribution. For the estimation of Δ_{log} and Δ_{wmw}, if we do not relate the ordinal responses to latent continuous variables, the likelihood function can be simply given by

$$\begin{aligned}
L &= \sum_{i=1}^{2} \sum_{k=1}^{K} n_{ik} \log(\pi_{ik}) \\
&= \sum_{i=1}^{2} \sum_{k=1}^{K} n_{ik} \log(\gamma_{ik} - \gamma_{i(k-1)})
\end{aligned} \tag{4.19}$$

where $\gamma_{i0} = 0$ and $\gamma_{iK} = 1$. The MLE for γ_{ik} are $\hat{\gamma}_{ik} = (n_{i1} + \cdots + n_{n_{ik}})/n_i$, $i = 1, 2$, $k = 1, \cdots, K-1$. By applying the relationship between Δ_{wmw} and γ_{ik}s as in (4.8), we can derive the MLE of Δ_{wmw} as

$$\hat{\Delta}_{wmw} = (n_1 n_2)^{-1} \sum_{k=2}^{K} n_{2k} \cdot \sum_{l=1}^{k-1} n_{1l} + 0.5 \cdot (n_1 n_2)^{-1} \sum_{k=1}^{K} n_{1k} n_{2k}.$$

This is the usual WMW test statistic (see e.g. Ryu and Agresti, 2008; Zhao, Rahardja, and Qu, 2008). If we consider the relationship between the common log-odds ratio Δ_{log} and the γ_{ik}s as in (4.5) under a proportional odds assumption, the MLE for Δ_{log} can be derived as (Dark, Bolland, and Whitehead, 2003)

$$\hat{\Delta}_{log} = Z/V,$$

where the detailed expression for Z and V will be given in Section 4.4.1.

If we specify a latent variable model and use (4.1) in the likelihood function (4.19), the distribution of treatment responses will have convenient location and scale interpretations, and the corresponding treatment effects can be expressed by the parameters of the underlying distributions. Based on our proposed estimation procedure, the estimates of γ_{ik}s in terms of the parameters of the latent variable model can be given by

$$\hat{\gamma}_{ik} = F_0(\frac{\hat{\tau}_k - \hat{\mu}_i}{\hat{\sigma}_i}), \quad i = 1, 2, \ k = 1, \cdots, K-1.$$

Thus, from (4.8) and (4.5), we can derive the estimates of Δ_{wmw} and Δ_{log}, respectively, based on the latent variable model. In fact, the common log-odds ratio Δ_{log} can be directly estimated based on the new interpretations on the proportional odds model. More specifically, we may specify the logistic distributions for the underlying random variables X_i, $i = 1, 2$, and assume them have equal scales. On the basis of the two-step estimation procedure, this equal scale assumption is equivalent to assume $X_1 \sim L(\mu_1, 1)$ and $X_2 \sim L(\mu_2, 1)$, where $L(\cdot, \cdot)$ denotes the logistic distribution. From (4.5) and the two-step estimation procedure, the the common log-odds ratio Δ_{log} can be directly derived as

$$\hat{\Delta}_{log} = \hat{\mu}_1 - \hat{\mu}_2.$$

So, this framework for the analysis of ordinal data via the latent variable model subsumes the existing methods. The validation of the existing methods usually depends on the equal scale assumption. This framework can give convenient interpretation and estimation of the location and scale of the treatments with ordinal responses. Moreover, some further statistical inference can be conducted based on such framework and will be detailed illustrated in the following subsection.

4.3.2 Statistical Inference

On the basis of the proposed latent variable model and the estimation procedure, many statistical inferences can be conducted conveniently.

(I). The goodness-of-fit test

To test the goodness-of-fit of the proposed model, the classical Pearson chi-square statistic that compares the observed cell counts and the model-based expected cell counts can be given by

$$\chi^2 = \sum_{i=1}^{G}\sum_{k=1}^{K} \frac{(n_{ik} - n_i\pi_{ik}(\hat{\boldsymbol{\theta}}))^2}{n_i\pi_{ik}(\hat{\boldsymbol{\theta}})}. \tag{4.20}$$

The deviance statistic also compares the observed cell counts and expected cell counts, but gives the test statistic in the form of a likelihood ratio test.

$$\begin{aligned} D^2 &= 2\sum_{i=1}^{G}\sum_{k=1}^{K} n_{ik}\log\frac{n_{ik}}{n_i\pi_{ik}(\hat{\boldsymbol{\theta}})} \\ &= -2\sum_{i=1}^{G}\sum_{k=1}^{K} n_{ik}\log\pi_{ik}(\hat{\boldsymbol{\theta}}) + 2\sum_{i=1}^{G}\sum_{k=1}^{K} n_{ik}\log\frac{n_{ik}}{n_i}. \end{aligned} \tag{4.21}$$

These test statistics are distributed as central chi-square with degrees of freedom df=$G(K-1)-(p-2)$, where p is the dimension of $\boldsymbol{\theta}$. Since in the two-step estimation procedure, the location and scale parameter of treatment s are actually fixed, $p-2$ can be regard as the effective dimension of $\boldsymbol{\theta}$. If the value of the test statistic exceeds the corresponding upper α critical value, the specified underlying distribution should be rejected. It should be noted that if $G(K-1) < p-2$, the model is not identifiable;

and if $G(K-1) = p-2$, the model is perfect fitted. Only when $G(K-1) > p-2$, the test of goodness-of-fit becomes sensible.

In the analysis framework, we use the underlying L-S distribution family to model the ordinal responses. Although other distributions are also the alternatives, the L-S family can usually provide adequate model fit, since the values of the thresholds, the locations, and the scales can be 'adjusted' in a very flexible manner to fit the ordinal data according to the maximum likelihood principle.

(II). The test of dispersion effect

When the underlying distributions are properly specified, the dispersion effect among the G treatments can be investigated by testing the null hypothesis (4.10). In this case, we may consider the following restricted log-likelihood function in the step 2 estimation procedure.

$$\tilde{L}_2(\boldsymbol{\theta}_1) = \sum_{i=1}^{G}\sum_{k=1}^{K} n_{ik}\log(\pi_{ik}(\boldsymbol{\theta}_1)), \tag{4.22}$$

where,

$$\pi_{ik}(\boldsymbol{\theta}_1) = F_0(\frac{\tau_k - \mu_i}{\sigma}) - F_0(\frac{\tau_{k-1} - \mu_i}{\sigma}).$$

Here, σ is the common scale parameter in (4.10), and $\boldsymbol{\theta}_1 = (\mu_1, \cdots, \mu_G, \sigma)'$ is the vector containing the unknown parameters. The score function and the Fisher information matrix for the likelihood function (4.22) are given in the Appendix D. So, it is easy to obtain the MLE of $\boldsymbol{\theta}_1$. Then, the following likelihood ratio test can be utilized to test (4.10), which will asymptotically have a chi-square distribution with degree of freedom df=$G-1$.

$$LRT = -2\tilde{L}_2(\hat{\boldsymbol{\theta}}_1) + 2L_2(\hat{\boldsymbol{\theta}}_0) = -2\sum_{i=1}^{G}\sum_{k=1}^{K} n_{ik}\log\frac{\pi_{ik}(\hat{\boldsymbol{\theta}}_1)}{\pi_{ik}(\hat{\boldsymbol{\theta}}_0)}, \tag{4.23}$$

where $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_0$ are the MLEs of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_0$ involved in (4.22) and (4.14), respectively.

(III). The test of location effect

To test the null hypothesis (4.11), a overall likelihood ratio test as in (II) can also be adopted. However, this overall test can not provide further information on

pairwise location effects between any two treatments. For this purpose, we may adopt the following multiple testing strategy.

From the asymptotic theory of maximum likelihood estimation and the analysis in Section 4.3.1, under suitable regularity conditions, for $1 \leq i \neq j \leq G$,

$$\frac{\hat{\mu}_i - \hat{\mu}_j - (\mu_i - \mu_j)}{\sqrt{\frac{\sigma_i^2}{n_i} \cdot \delta(\mu_i, \sigma_i) + \frac{\sigma_j^2}{n_j} \cdot \delta(\mu_j, \sigma_j)}} \xrightarrow{L} N(0,1)$$

So, we may use the following statistic to test the location effect between treatment i and j.

$$Z_{ij} = \frac{\hat{\mu}_i - \hat{\mu}_j}{\sqrt{\frac{\hat{\sigma}_i^2}{n_i} \cdot \hat{\delta}(\hat{\mu}_i, \hat{\sigma}_i) + \frac{\hat{\sigma}_j^2}{n_j} \cdot \hat{\delta}(\hat{\mu}_j, \hat{\sigma}_j)}}, \tag{4.24}$$

where $\hat{\delta}(\hat{\mu}_i, \hat{\sigma}_i)$ with a hat denotes that the value of $\delta(\cdot, \cdot)$ is evaluated based on the estimates of $\boldsymbol{\tau}$, μ_i, and σ_i. From the consistency property of MLEs and Slutsky's theorem, Z_{ij} will also converges in distribution to $N(0,1)$ under the null hypothesis (4.11).

The null hypothesis (4.11) can be tested by some multiple testing procedures based on the statistics given in (4.24). For example, the multiple testing of several treatments with a control can be conducted as in Chapter 3. In addition, the pairwise comparison can also be conducted by testing all pairwise treatment effects. In these multiple testing procedures, the correlations among the test statistics are usually very informative. They can be derived conveniently. Specifically, for all unequal i, i', j, and j',

$$Corr(Z_{ij}, Z_{i'j}) = 1/\sqrt{(\frac{n_j}{n_i}\frac{\sigma_i^2\delta(\mu_i, \sigma_i)}{\sigma_j^2\delta(\mu_j, \sigma_j)} + 1)(\frac{n_j}{n_{i'}}\frac{\sigma_{i'}^2\delta(\mu_{i'}, \sigma_{i'})}{\sigma_j^2\delta(\mu_j, \sigma_j)} + 1)},$$

and $Corr(Z_{ij}, Z_{i'j'}) = 0$.

(IV). Model selection

The underlying continuous distributions in our discussion are limited to the L-S family. A nature question is how to select the most suitable distribution from this family for a given ordinal data set. For this issue, we may use the model selection

criteria, such as AIC and BIC, to determine the suitable underlying distribution. This two criteria can be given by

$$AIC = -2\text{log-likelihood}_{max}(\boldsymbol{\theta}) + 2\dim(\boldsymbol{\theta})$$
$$BIC = -2\text{log-likelihood}_{max}(\boldsymbol{\theta}) + \log N \cdot \dim(\boldsymbol{\theta}),$$

where $\dim(\boldsymbol{\theta})$ denotes the number of parameters in $\boldsymbol{\theta}$, and N is the total sample size.

The general idea behind these criteria is to penalize the maximum log-likelihood function by the complexity of the model. The BIC clearly has a stronger penalty for complexity. Since our model has the same complexity (the same number of unknown parameters) for different underlying distributions, The usage of AIC or BIC as a model selection criterion is equivalent to the use of the deviance statistic (4.21). Note that the second term in (4.21) is corresponding to the log-likelihood function of the saturated model. So, we suggest to use the deviance statistic (4.21) as a criterion to select the proper underlying distribution. For example, if the deviance statistic for the latent logistic distribution model is larger than that for the latent normal distribution model, it is more appropriate to use the latent normal distribution to model the observed ordinal data.

At the end of this section, we want to point out that the testing results of the treatment effects (the dispersion effect and location effect) based on test statistics (4.23) and (4.24) will not be influenced by the selection of different treatment to determine the thresholds in step 1. This result can be summarized in the following lemma. The proof of the Lemma 4.1 is given in Appendix C.

Lemma 4.1 When sample size is large, the different selection of treatment s in step 1 of the two-step estimation procedure has no effect on the testing results.

Treatment	C_1	C_2	$\cdots$	C_K	Total
Treatment 1	n_{11}	n_{12}	$\cdots$	n_{1K}	n_1
Treatment 2	n_{21}	n_{22}	$\cdots$	n_{2K}	n_2
Total	S_1	S_2	$\cdots$	S_K	N

Table 4.1: Ordinal data of two treatments organized in contingency table

4.4 The Power and Sample Size Determination in the Comparison of Two Independent Treatments

In this section, we consider the comparison of two treatments with ordinal responses by utilizing the proposed latent variable method. The corresponding sample size determination based on the latent variable method is also proposed. This newly developed method is compared with the existing methods in terms of power and sample size determination.

We first outline two existing testing and sample size determination methods for the comparison of two treatments with ordered categorical responses. The ordered categorical observations can be organized in Table 4.1.

4.4.1 The Existing Methods

1. The Wilcoxon-Mann-Whitney test (WMW)

The WMW test quantifies the treatment effect by the competing probability given in (4.7). The null hypothesis that indicates there is no difference between the two treatments is given by

$$H_0 : \Delta_{wmw} = 0.5$$

The WMW tests the null hypothesis by constructing a z-statistic

$$z_0 = \frac{\hat{\Delta}_{wmw} - 0.5}{\hat{\sigma}_{H_0}(\hat{\Delta}_{wmw})} \sim N(0,1),$$

where

$$\hat{\Delta}_{wmw} = (n_1 n_2)^{-1} \sum_{k=2}^{K} n_{2k} \cdot \sum_{l=1}^{k-1} n_{1l} + 0.5 \cdot (n_1 n_2)^{-1} \sum_{k=1}^{K} n_{1k} n_{2k},$$

and

$$\hat{\sigma}^2_{H_0}(\hat{\Delta}_{wmw}) = \frac{N+1}{12 n_1 n_2} - \frac{1}{12N(N-1)n_1 n_2} \sum_{k=1}^{K} (S_k^3 - S_k)$$

is the variance of $\hat{\Delta}_{wmw}$ under the null hypothesis (Lehmann, 1975, p.20), where $S_k = n_{1k} + n_{2k}$, $k = 1, \cdots, K$.

We only consider the two-sided alternative hypothesis. The null hypothesis will be rejected with significance level α, if

$$|\hat{\Delta}_{wmw} - 0.5| > z_{\alpha/2} \cdot \hat{\sigma}_{H_0}(\hat{\Delta}_{wmw})$$

where $z_{\alpha/2}$ is the upper $\frac{\alpha}{2}$-level significance point of standard normal distribution,

To consider the sample size determination issue, let

$$n_1 = Nt, \quad n_2 = N(1-t), \quad p_k = \frac{n_{2k}}{n_2} = \frac{n_{2k}}{N(1-t)}, \quad q_k = \frac{n_{1k}}{n_1} = \frac{n_{1k}}{Nt}, \text{ and}$$

$$S_k = n_{1k} + n_{2k} = N[tq_k + (1-t)p_k].$$

For given significance level α and power $1-\beta$, the sample size N for two-sided hypothesis can be obtained by solving the following equation

$$\left(\frac{\hat{\Delta}_{wmw} - 0.5}{\sigma_{H_0}(\hat{\Delta}_{wmw})} \right)^2 = \left(z_{\alpha/2} + \frac{\sigma_{H_a}(\hat{\Delta}_{wmw})}{\sigma_{H_0}(\hat{\Delta}_{wmw})} z_\beta \right)^2,$$

where $\sigma_{H_a}(\hat{\Delta}_{wmw})$ is the standard deviation of $\hat{\Delta}_{wmw}$ under the alternative hypothesis. Zhao, Rahardja, and Qu (2008) used the assumption that $\sigma_{H_a}(\hat{\Delta}_{wmw}) = \sigma_{H_0}(\hat{\Delta}_{wmw})$, and derived the sample size formula for the WMW test as follows.

$$N_{wmw} = \frac{(z_{\alpha/2} + z_\beta)^2 (1 - \sum_{k=1}^{K} ((1-t)p_k + tq_k)^3)}{12t(1-t)(\hat{\Delta}_{wmw})^2}, \tag{4.25}$$

where $\hat{\Delta}_{wmw} = \sum_{k=2}^{K} p_k \sum_{l=1}^{k-1} q_l + 0.5 \sum_{k=1}^{K} p_k q_k$, which is expressed in terms of the cell proportions.

2. Whitehead (1993) testing method (Whd)

Whitehead (1993) proposed a method to measure the treatment effect in terms of log-odds-ratio based on the proportional odds assumption. That is, (4.5) holds true for $k = 1, \cdots, K-1$. Here we denote the treatment effect measure based on Whitehead (1993) method by Δ_{whd}. Specifically, the common log-odds ratio is defined by

$$\Delta_{whd} = \log\left(\frac{\gamma_{2k}(1-\gamma_{1k})}{\gamma_{1k}(1-\gamma_{2k})}\right) \quad \text{for all } k.$$

The null hypothesis of no treatment effect is given by

$$H_0 : \Delta_{whd} = 0.$$

Whitehead (1993) gave the test statistic Z that follows approximately normal distribution with mean $\Delta_{whd}V$ and variance V, where

$$Z = \frac{1}{N+1}\sum_{k=1}^{K} n_{1k}(L_{2k} - U_{2k}),$$

and

$$V = \frac{n_1 n_2 N}{3(N+1)^2}\left[1 - \sum_{k=1}^{K}\left(\frac{S_k}{N}\right)^3\right].$$

In the expression of Z, L_{2k} and U_{2k} are the lower and upper cumulative totals of treatment 2, which can be calculated by

$$L_{2k} = n_{21} + \cdots + n_{2(k-1)}, \qquad \text{for } k = 2, \cdots, K, \quad \text{and}$$

$$U_{2k} = n_{2(k+1)} + \cdots + n_{2K}, \qquad \text{for } k = 1, \cdots, K-1,$$

with $L_{21} = U_{2K} = 0$.

For a given significance level α, the null hypothesis will be rejected if

$$|Z| > \sqrt{V} \cdot z_{\alpha/2}.$$

Whitehead (1993) pointed out that this test statistic is essentially the MWM test. So, the power of this test should be identical with the power of the WMW test. Our simulation study (see Table 4.2 in Section 4.4.3) further verifies this argument.

Dark, Bolland, and Whitehead (2003) derived that the MLE of the common log-odds ratio Δ_{whd} is Z/V.

The power of the test against the alternative $H_a : \Delta_{whd} = \Delta^a_{whd} \neq 0$ is given by

$$\text{power} = P(|Z| > \sqrt{V} \cdot z_{\alpha/2} | H_a) = 1 - \beta.$$

This reduces to

$$V = \left[\frac{z_{\alpha/2} + z_\beta}{\Delta^a_{whd}} \right]^2.$$

We still suppose $n_1 = N \cdot t$ and $n_2 = N \cdot (1-t)$, $0 < t < 1$. Let $\bar{p}_k = \frac{S_k}{N}$. Then, the sample size formula can be given by

$$N_{whd} = \frac{3 \cdot (z_{\alpha/2} + z_\beta)^2}{t \cdot (1-t) \cdot (\Delta^a_{whd})^2 \cdot (1 - \sum_{k=1}^{K} \bar{p}_k^3)}. \tag{4.26}$$

Theoretically, the sample size formula (4.26) should produce the same result as (4.25) for given data set, since they are essentially derived from the same test statistic. Our calculations (see Table 4.3 in Section 4.4.3, and Table 4.6 in Section 4.6) also verify this point.

4.4.2 The Latent Variable Method

In this subsection, we will give a detailed illustration of the latent variable method in the comparison of two treatments with ordinal responses. The corresponding sample size determination method based on the latent variable model is also presented. We focus on the detection of the location effect between two treatments while accommodate the heterogeneous scales in the treatments.

From the arguments in Section 4.3.1 and the Appendix C, the comparison of treatments is actually in a relative sense on the basis of the analysis framework. Without loss of generality, we may assume the true underlying distributions of the two treatments are $F(x; \mu_1, \sigma_1) = F(x; 0, 1)$ for treatment 1 and $F(x; \mu_2, \sigma_2)$ for treatment 2. We will discuss two important latent variable models, the latent normal distribution model and the latent logistic distribution model, and denote this two

methods by *LNorm* and *LLogis*, respectively. The corresponding quantities for the two methods are distinguished by adding subscript *norm* and *logis*, respectively.

Thus, the null hypothesis of no location effect can be given by

$$H_0 : \Delta_{norm} = \mu_2 - \mu_1 = 0 \quad \text{if } F(x; \mu_i, \sigma_i) = N(\mu_i, \sigma_i), \quad i = 1, 2,$$

or,

$$H_0 : \Delta_{logis} = \mu_2 - \mu_1 = 0 \quad \text{if } F(x; \mu_i, \sigma_i) = L(\mu_i, \sigma_i), \quad i = 1, 2,$$

where $N(\cdot, \cdot)$ and $L(\cdot, \cdot)$ denote the normal distribution function and the logistic distribution function respectively.

From (4.24), we can test the null hypothesis by constructing the following z-statistic

$$Z_{21} = \frac{\hat{\mu}_2 - \hat{\mu}_1}{\hat{\sigma}_{\hat{\mu}_2 - \hat{\mu}_1}} = \frac{\hat{\mu}_2 - \hat{\mu}_1}{\sqrt{\frac{\hat{\sigma}_2^2}{n_2}\hat{\delta}(\hat{\mu}_2, \hat{\sigma}_2) + \frac{1}{n_1}\hat{\delta}(0, 1)}} \sim N(0, 1). \tag{4.27}$$

Thus, the null hypothesis will be rejected with significance level α, if

$$|\hat{\mu}_2 - \hat{\mu}_1| > z_{\alpha/2}\hat{\sigma}_{\hat{\mu}_2 - \hat{\mu}_1}.$$

If the alternative hypothesis is specified as $H_a : \Delta^a = \mu_2 - \mu_1 = \Delta^a_{norm} \neq 0$ for the *LNorm* method, and $H_a : \Delta^a = \mu_2 - \mu_1 = \Delta^a_{logis} \neq 0$ for the *LLogis* method. Then, the power of the test against the alternative is given by

$$\text{power} = P(|\hat{\mu}_2 - \hat{\mu}_1| > z_{\alpha/2}\hat{\sigma}_{\hat{\mu}_2 - \hat{\mu}_1} | H_a) = 1 - \beta.$$

This reduces to

$$\left(z_{\alpha/2} + z_\beta\right)^2 = \left(\frac{\Delta^a}{\hat{\sigma}_{\hat{\mu}_2 - \hat{\mu}_1}}\right)^2.$$

We still denote $n_1 = Nt$, $n_2 = N(1 - t)$, $0 < t < 1$. Then, the required sample size to detect a significant location effect equal to Δ^a with power $1 - \beta$ can be calculated by

$$N_{norm} = \frac{(z_{\alpha/2} + z_\beta)^2[t\hat{\sigma}_2^2\hat{\delta}(\hat{\mu}_2, \hat{\sigma}_2) + (1 - t)\hat{\delta}(0, 1)]}{t(1 - t)(\Delta^a_{norm})^2}, \quad \text{if } LNorm; \tag{4.28}$$

$$N_{logis} = \frac{(z_{\alpha/2} + z_{\beta})^2[t\hat{\sigma}_2^2\hat{\delta}(\hat{\mu}_2, \hat{\sigma}_2) + (1-t)\hat{\delta}(0,1)]}{t(1-t)(\Delta^a_{logis})^2}, \quad \text{if } LLogis. \tag{4.29}$$

A special case of the above discussion is that the two treatments have equal scales. For this special case, the treatment effect is completely attributed to location effect, and it is equivalent to assume underlying distributions are $F(x;\mu_1,1)$ for treatment 1 and $F(x;\mu_2,1)$ for treatment 2 under the analysis framework. Thus, for this equal scale case, the calculation of N_{norm} and N_{logis} can also be performed by (4.28) and (4.29), respectively, only with the replacement of $\hat{\sigma}_2$ by 1 in the formulae.

The implementation of formulae (4.28) and (4.29) depends on a prior study of the two treatments and the derivation of the estimates of τ, μ_2, and σ_2. Usually, in the design stage of an experiment, it is difficult to derive the estimates of μ_2 and σ_2. For this case, we may take the equal scale assumption, and let $\mu_2 = \Delta^a$ by noticing that $\Delta^a = \mu_2 - \mu_1$ with $\mu_1 = 0$ under the analysis framework. Therefore, we propose the following sample size determination formulae for the case no prior information available.

$$\tilde{N}_{norm} = \frac{(z_{\alpha/2} + z_{\beta})^2[t\hat{\delta}(\Delta^a_{norm}, 1) + (1-t)\hat{\delta}(0,1)]}{t(1-t)(\Delta^a_{norm})^2}, \quad \text{if } LNorm; \tag{4.30}$$

$$\tilde{N}_{logis} = \frac{(z_{\alpha/2} + z_{\beta})^2[t\hat{\delta}(\Delta^a_{logis}, 1) + (1-t)\hat{\delta}(0,1)]}{t(1-t)(\Delta^a_{logis})^2}, \quad \text{if } LLogis. \tag{4.31}$$

The sample size formulae (4.30) and (4.31) only require the cell proportions of treatment 1 (to determine the thresholds), which usually serves as a reference or control treatment and it is relatively easy to derive these cell proportions. However, the determination of N_{wmw} and N_{whd} depends on the cell proportions of both treatments, which is not so convenient to implement.

The sample size formulae (4.28) and (4.29) are derived by specifying treatment 1 as a reference with underlying distribution $F(x;\mu_1,\sigma_1) = F(x;0,1)$. The sample size formulae that were derived by specifying treatment 2 as a reference should produce the same results as (4.28) and (4.29), respectively. A theoretical proof can be simply worked out by following a similar argument as in Appendix C. Since it is a intuitive result (also verified by calculation), we do not present the details here.

Finally, we want to point out that, when the underlying variables have equal scales, this newly developed *LLogis* method is actually the Whd method, which is derived based on an alternative interpretation of the proportional odds model. Specifically, when the two treatments have equal scales, $\Delta_{logis} = -\Delta_{whd}$, and these two methods have the same power of test (see Table 4.2) and almost the same sample size results (See Table 4.3). A combination with the discussion at the end of Section 4.4.1 implies that these three methods, the WMW, the Whd, and the *LLogis* method are equivalent in terms of power and sample size determination, when the treatments have equal scales.

4.4.3 The Comparison of Different Methods

1. Power comparison: a simulation study

A simulation study is conducted to compare the power of the four testing methods, WMW, Whd, *LNorm*, and *LLogis*. The underlying distributions, $F(x; \mu_i, \sigma)$, are assumed to have equal scale parameters for the two treatments and belong to the L-S family. The true underlying distributions (TUD) chosen for the simulation are normal ($N(\mu_i, \sigma)$ with mean μ_i and standard deviation σ), logistic ($L(\mu_i, \sigma)$ with mean μ_i and scale σ), and Cauchy ($C(\mu_i, \sigma)$ with location μ_i and scale σ). The thought behind these choices is to include average-tailed and heavy-tailed distributions. When continuous deviates are generated from theses distributions, we can derive the ordinal data for different treatments based on the latent variable model and a set of fixed thresholds. In our simulation, the true thresholds used to generate ordinal data are set as (-1.5,-0.5,0.5,1.5) for all three types of true underlying distributions. We assume the two treatments have equal sample size $n_1 = n_2 = n$. The nominal significant level is fixed at $\alpha = 0.05$. The powers of different testing methods are presented in Table 4.2, including the estimates of the type-I errors (marked with asterisk). All these simulation results are based on 100,000 replications.

From the simulation results, we find that the four testing methods can control

TUD		Size	Power (Type-I error*)			
$F(\mu_1, \sigma)$	$F(\mu_2, \sigma)$	n	LNorm	LLogis	Whd	WMW
N(0,1)	N(0,1)	1000	0.0498*	0.0499*	0.0495*	0.0495*
N(0.2,2)	N(0.2,2)	1000	0.0508*	0.0514*	0.0511*	0.0511*
N(0,1)	N(0.1,1)	1000	0.5348	0.5195	0.5180	0.5180
N(0,1)	N(0.2,1)	500	0.8460	0.8346	0.8331	0.8330
N(0,1)	N(0.2,1)	200	0.4654	0.4526	0.4501	0.4489
N(0.1,2)	N(0.4,2)	500	0.6008	0.5997	0.5996	0.5996
L(0,1)	L(0,1)	500	0.0472*	0.0481*	0.0478*	0.0477*
L(0.1,2)	L(0.1,2)	500	0.0485*	0.0482*	0.0493*	0.0492*
L(0,1)	L(0.1,1)	1000	0.2289	0.2353	0.2347	0.2345
L(0,1)	L(0.2,1)	500	0.4178	0.4307	0.4300	0.4298
L(0,1)	L(0.3,1)	500	0.7312	0.7442	0.7448	0.7447
L(0.1,2)	L(0.4,2)	500	0.2596	0.2603	0.2605	0.2604
C(0,1)	C(0,1)	500	0.0470*	0.0467*	0.0467*	0.0464*
C(0.2,2)	C(0.2,2)	500	0.0460*	0.0458*	0.0465*	0.0465*
C(0,1)	C(0.1,1)	1000	0.1884	0.2090	0.2086	0.2085
C(0,1)	C(0.2,1)	500	0.3326	0.3800	0.3801	0.3792
C(0,1)	C(0.3,1)	500	0.6421	0.7097	0.7102	0.7097
C(0.1,2)	C(0.4,2)	500	0.2806	0.2916	0.2910	0.2908

Table 4.2: The power of different testing methods.

the type-I error at 5%. The simulation results on the power verify our argument that the three methods, WMW, Whd, and *LLogis* have almost the same power. Another important yet intuitive finding is that the *LNorm* method will be more powerful when the underlying variables have average-tailed distributions, and the other three methods will be more powerful when the underlying variables have heavy-tailed distributions. This finding can serve as a rough criterion in selecting the underlying distribution for the model by observing the distributions of the ordinal responses.

2. Sample size comparison: an accurate calculation

The sample size formulae of different methods are function of the significance level α, allocation ratio t, power $1-\beta$, and the treatment effect Δ. In this section, we conduct a calculation to study the sample size determination of different methods, where the treatment effect is measured by the location parameters of the underlying

distributions. The underlying distributions are still assumed to have equal scales. Without loss of generality, we assume this identical scale is 1.

Since both the *LLogis* method and the Whd method depend on the underlying logistic distribution assumption and the *LNorm* method depends on the underlying normal distribution assumption, while the WMW test does not depend on a specific form of the underlying distribution, here we only consider the usage of logistic distribution and normal distribution as the underlying distributions. Specifically, we set $N(0,1)$ and $N(\Delta,1)$ as the underlying normal distributions, and $L(0,1)$ and $L(\Delta,1)$ as the underlying logistic distributions. Thus, the common treatment effects are measured by Δ. When the values of the thresholds (set as (-1.5,-0.5,0.5,1.5) in our calculation) are specified, the cumulative probabilities γ_{ik}s can be calculated based on the underlying distributions. Then, we may calculate the different treatment effect measures, Δ_{norm}, Δ_{logis}, Δ_{whd}, and Δ_{wmw}, and the resulting sample size of different methods.

When the underlying distribution is specified as normal distribution, we only consider the sample size based on the *LNorm* method and the WMW method; and when the underlying distribution is logistic distribution, we only calculate the sample size base on the *LLogis* method, the Whd method, and the WMW method. Because only in these cases the corresponding treatment effects can be derived exactly. The calculation results are presented in Table 4.3.

From the calculation results in Table 4.3, when the underlying distribution is specified as logistic, these three methods, *LLogis*, Whd, and WMW produce very close sample size results. The sample sizes derived by *LLogis* and Whd are almost the same, and they are a little less than the result derived by the WMW method. When the true underlying distribution is normal, the required sample size of the *LNorm* method is also a little less than that of the WMW method. This means that the latent variable method is more efficient that the WMW method. The same result can also be drawn from the power comparison in Table 4.2.

			TUD: $N(0,1)$, $N(\Delta,1)$				TUD: $L(0,1)$, $L(\Delta,1)$					
Effect	Ratio	Power	LNorm		WMW		LLogis		Whd		WMW	
Δ	t	$1-\beta$	Δ_{norm}	N_{norm}	Δ_{wmw}	N_{wmw}	Δ_{logis}	N_{logis}	Δ_{whd}	N_{whd}	Δ_{wmw}	N_{wmw}
0.1	0.5	0.8	0.1	3444	0.526	3551	0.1	9830	-0.1	9829	0.516	9835
0.2	0.5	0.8	0.2	862	0.552	892	0.2	2459	-0.2	2458	0.532	2464
0.3	0.5	0.8	0.3	384	0.577	400	0.3	1094	-0.3	1093	0.548	1099
0.4	0.5	0.8	0.4	216	0.603	228	0.4	617	-0.4	615	0.563	621
0.5	0.5	0.8	0.5	139	0.628	148	0.5	395	-0.5	394	0.579	400
0.8	0.5	0.8	0.8	55	0.698	61	0.5	156	-0.5	154	0.625	161
0.5	0.2	0.8	0.5	216	0.628	230	0.5	616	-0.5	617	0.579	623
0.5	0.8	0.8	0.5	218	0.628	231	0.5	620	-0.5	614	0.579	626
0.5	0.5	0.6	0.5	87	0.628	92	0.5	247	-0.5	246	0.579	250
0.8	0.5	0.6	0.8	34	0.698	38	0.8	98	-0.8	96	0.625	100

Table 4.3: The sample size determination of different methods.

	Retinopathy status			
Group	None	Non-proliferative	Advanced	Total
Non-smoking	191 (66.32%)	42 (14.58%)	55 (19.10%)	288 (100%)
Smoking	197 (60.62%)	76 (23.38%)	52 (16.00%)	325 (100%)
Total	388	118	107	613

Table 4.4: Retinopathy category by smoking status for 613 diabetic patients.

4.5 A Real Data Example

We consider a six-year follow-up study by Bender and Grouven (1998) where 613 type-1 diabetic patients were studied for associations between retinopathy status (RS) and smoking (SM), adjusted for the known risk factors, diabetes duration (DD), glycosylated hemoglobin (H1C), and diastotic blood pressure (DBP). Ratinopathy status is defined by three ordered categories: 0=no retinopathy; 1=non-proliferative retinopathy; and 2=advanced retinopathy or blind. Smoking status is a binary variable which equals to 1 if the patient smoked during the study and 0 otherwise.

This real data example is also studied by Rabbee, Coull, and Mehta (2003) and Zhao, Rahardja, and Qu (2008), respectively, where both of them focus on only the effect of smoking on the retinopathy status without controlling for the other covariates. The data set they considered are displayed in Table 4.4. Rabbee *et al.* (2003) use an approximate and improved power function of Whitehead's (1993) test statistic to study the example, and show that the power for the given sample size is too low to detect a significant treatment difference. Rabbee *et al.* (2003) suggested that a larger sample size should be used in the planning stage of this study, while they did not give a clear sample size formula for their method. Zhao *et al.* (2008) studied the power and required sample size for this example using the WMW test. The WMW test on the data is not significant with a p-value > 0.3. The total sample size for some different alternatives of the WMW test based on nominal power requirement are also calculated in their paper.

In this section, we give a study of the retinopathy data using our proposed latent

variable method. For this data set with two treatments and three categories, the model is perfect fitted. It is not sensible to consider the model selection by the deviance statistic (4.21). Thus, in this section, we give the analysis results based on both the *LNorm* method and the *LLogis* method. We add subscript *norm* or *logis* to the quantities to distinguish the estimation results of the two methods.

We consider the Non-smoking group as treatment 1 and the Smoking group as treatment 2 in the estimation. From the step 1 procedure (with $s = 1$), we get the estimates of the thresholds, $\hat{\boldsymbol{\tau}}_{norm} = (0.4212, 0.8743)$ and $\hat{\boldsymbol{\tau}}_{logis} = (0.6776, 1.4437)$. From the step 2 procedure, we obtain that $\hat{\boldsymbol{\theta}}_{0,norm} = (\hat{\mu}_1, \hat{\sigma}_1, \hat{\mu}_2, \hat{\sigma}_2) = (0.0, 1.0, 0.2529, 0.6249)$ and $\hat{\boldsymbol{\theta}}_{0,logis} = (\hat{\mu}_1, \hat{\sigma}_1, \hat{\mu}_2, \hat{\sigma}_2) = (0.0, 1.0, 0.4083, 0.6244)$. The values of the -2 times log-likelihood are given by $-2 \cdot L_2(\hat{\boldsymbol{\theta}}_{0,norm})$=1109.43 and $-2 \cdot L_2(\hat{\boldsymbol{\theta}}_{0,logis}) = 1109.43$.

We conduct the testing of the following two null hypotheses

$$H_{01} : \sigma_2 = \sigma_1 = \sigma, \quad \text{and,} \quad H_{02} : \mu_2 - \mu_1 = 0.$$

The likelihood ratio test (4.23) can be used to test H_{01}. The values of the -2 times log-likelihood with the constraint $\sigma_2 = \sigma_1 = \sigma$ are given by $-2 \cdot \tilde{L}_2(\hat{\boldsymbol{\theta}}_{1,norm}) = 1116.89$, and $-2 \cdot \tilde{L}_2(\hat{\boldsymbol{\theta}}_{1,logis}) = 1116.51$. Thus, we have $LRT_{norm} = 7.46$ and $LRT_{logis} = 7.08$. The comparison of these values with the critical value of the chi-squared distribution with 1 degree of freedom (the corresponding p-values less than 0.01) implies that the null hypothesis H_{01} should be rejected with high significant level. This phenomenon of unequal dispersions of the two groups has not been accommodated or observed in any of the analysis in the literature.

Since this two treatments have significant dispersion effect, we use the test statistic (4.27) that involves the dispersion information for the test of H_{02}. After calculation, we derive that $Z_{norm} = 2.079$ and $Z_{logis} = 2.119$ with the corresponding p-values $p_{norm} = 0.038$ and $p_{logis} = 0.034$, respectively. However, if we use formula (4.27) under the equal scale assumption (replace $\hat{\sigma}_2$ by 1 in the formula), the corresponding results are $Z_{norm} = 1.868$ with $p_{norm} = 0.062$, and $Z_{logis} = 1.904$

Effect	N_{norm}		$\tilde{N}_{norm}$		Effect	N_{logis}		$\tilde{N}_{logis}$	
Δ_{norm}	t_1	t_2	t_1	t_2	Δ_{logis}	t_1	t_2	t_1	t_2
0.15	3165	3045	4220	4164	0.2	4468	4297	6081	6008
0.20	1780	1712	2285	2248	0.3	1986	1910	2573	2532
0.2529	*1113*	1071	1379	1353	0.4083	*1072*	1031	1328	1303
0.30	791	761	953	933	0.45	882	849	1077	1055
0.35	581	559	683	667	0.50	715	688	858	840
0.40	445	428	512	499	0.60	496	477	580	566
0.50	285	274	318	309	0.70	365	351	417	406
0.80	111	107	124	121	1.0	179	172	198	192

Table 4.5: The determination of sample size for the real example

with $p_{logis} = 0.057$. Different conclusions will be drawn if we take significant level $\alpha = 0.05$. This means that the dispersion information is very useful in the testing of the location effect. In other words, neglecting the dispersion difference in the two groups and simply comparing the location difference will lead to misleading results.

Now, we consider the sample size determination based on the *LNorm* method and the *LLogis* method. Since the equal scale assumption has been rejected, we use the formulae (4.28) and (4.29) that utilize the information from the prior study. As a comparison, the results based on formulae (4.30) and (4.31) are also presented. In the calculation, the observed proportions of the Non-smoking treatment, (66.32%, 14.58%, 19.10%), are used to determine the thresholds. Table 4.5 presents the required sample size for the two methods to detect a significant location effect Δ with significance level $\alpha = 0.05$ and power $1 - \beta = 0.8$. For each case, we consider two different allocation ratios, $t_1 = 0.4698$ (288/613) from the data and $t_2 = 0.5$ for balance allocation.

For the results presented in Table 4.5, it is interesting to notice the sample size highlighted by italic type, which are calculated based on the full information obtained from the observed samples. To achieve a power equal to 80%, the required sample size for the *LNorm* method is 1113, and the required sample size for the *LLogis* method is 1072. From the testing and sample size results of the this example,

it is more suitable to adopt the *LLogis* method for this data set. This sample size is much less than the sample size given by Zhao *et al.* (2008), where the required sample size is N_{wmw}=8390 from formula (4.25) that is also calculated based on the observed sample. Note that in Zhao's calculation, the observed proportions in smokers and non-smokers are only accurate to two places of decimals. If we correct these proportions to 4 decimal places as displayed in Table 4.4, we are surprised to find that the required sample size becomes N_{wmw}=6058 with $\hat{\Delta}_{wmw}$=0.5178. This finding indicates that the sample size formula of the WMW test is very sensitive to the slight changes of the cell proportions for this example.

In the calculation for this example, if we take the Smoking group as a reference (to determine the thresholds in step 1), we will derive the same testing and sample size results as what have been presented. If we apply the Whd method to the real data in Table 4.4, we derive that $\hat{\Delta}_{whd} = -0.1462$ and $N_{whd} = 6002$.

4.6 Further Study on Sample Size Determination

From the study of the example in Section 4.5, the determined sample sizes by the latent variable method and the WMW method (or Whd method) are markedly different. The reasons may be attributed to the significant difference in the dispersion effect between the two treatments. In this section, further numerical study is conducted to examine the behavior of the sample size formulae of different methods when the two treatments have different scales.

The required sample sizes for different methods are calculated when the ordinal responses of the two treatments are given. The ordered categorical data of the two treatments are generated based on the latent variable model with given specification of the true parameters. The specification of the true parameters includes $F(x;\mu_1,\sigma_1)$, $F(x;\mu_2,\sigma_2)$, $(\tau_1,\cdots,\tau_{K-1})$, the total sample size N, and the allocation ratio t. More specifically, the cell count n_{ik} of treatment i, $i = 1, 2$, in category C_k,

$k = 1, \cdots, K$, is derived by

$$n_{1k} = round(N \cdot t \cdot [F(\tau_k; \mu_1, \sigma_1) - F(\tau_{k-1}; \mu_1, \sigma_1)]),$$
$$n_{2k} = round(N \cdot (1-t) \cdot [F(\tau_k; \mu_2, \sigma_2) - F(\tau_{k-1}; \mu_2, \sigma_2)]),$$

where $round(\cdot)$ denotes the round off function. In our calculation, we fix $N = 600$ and $t = 0.5$, and different choices of other parameters are considered.

For given ordinal responses of the two treatments, the estimates of the treatment effects of different methods, $\hat{\Delta}_{wmw}$, $\hat{\Delta}_{whd}$, $\hat{\Delta}_{norm}$, and $\hat{\Delta}_{logis}$ are first calculated. Based on the estimation results of the treatment effects, the required sample sizes, N_{wmw}, N_{whd}, N_{norm}, and N_{logis} are calculated by formulae (4.25), (4.26), (4.28), and (4.29), respectively, with $\alpha = 0.05$, $1 - \beta = 0.8$, and Δ^a being assigned at the values of the estimated treatment effects. The calculation results are reported in Table 4.6.

From the results in Table 4.6, some important findings are summarized as follows.

(a) Consider the case that the center of the thresholds depart from zero to a side and $\sigma_1 > \sigma_2$ as indicated by case (A) and (C). In this case, as the true location difference $\mu_2 - \mu_1$ increases, the sample size N_{whd} and N_{wmw} increase very fast, reaching extremely large value for some larger location difference. This finding obviously contradicts our basic knowledge that it should require less sample size when the true location effect to be detected increases.

(b) Consider the case that the center of the thresholds depart from zero to a side and $\sigma_1 < \sigma_2$ as indicated by case (B) and (D). In this case, the the calculated sample sizes for N_{whd} and N_{wmw} become too small. We can believe that the calculated sample size N_{whd} and N_{wmw} are also questionable for this case.

(c) Comparing the results in case (D) and (E) indicates that the position of the thresholds will have a significant influence on the sample size N_{whd} and N_{wmw} when $\sigma_1 \neq \sigma_2$. The calculation in case (F) further substantiates this argument. This means that the skewness of the ordinal responses also have a significant effect on the sample size formulae N_{whd} and N_{wmw}.

(d) When the underlying distributions $F(x; \mu_i, \sigma_i)$ are normal, the estimated treatment effect $\hat{\Delta}_{norm}$ is very close to the true mean difference $(\mu_2 - \mu_1)/\sigma_1$; when the underlying distributions are specified as logistic distributions $L(\mu_i, \sigma_i)$, the estimated treatment effect $\hat{\Delta}_{logis}$ is also very close to the true mean difference $(\mu_2 - \mu_1)/\sigma_1$. From case (G), when the underlying distributions are logistic and $\sigma_1 = \sigma_2 = \sigma$, the estimated treatment effect $\hat{\Delta}_{logis}$ is very close to $-\hat{\Delta}_{whd}$, and they are both close to the true mean difference $(\mu_2 - \mu_1)/\sigma$ as indicated in (4.5). In this case, the four sample size determination methods have similar results.

In all these calculations, if we exchange the labels of treatment 1 and treatment 2, that is, taking treatment 2 as reference, it is obvious that all the values of the treatment effect measures will be different from what are presented in Table 4.6. However, the corresponding calculated sample sizes are the same.

Obviously, the ordinal data sets in cases (A) to (F) violate the proportional odds assumption. The score tests (SAS Proc Logistic) are significant for cases (A) to (E) with very small p-values. However, the score tests for case (F) are not significant with p-values larger than 0.3. The likelihood ratio test (4.23) based on *LLogis* method give consistent testing results with the score test.

In conclusion, the numerical study indicates that the sample size determination methods based on the Whd method and the WMW method will be quite questionable when the two treatments have different scale parameters, especially for the case that the center of the thresholds departs far from zero. The problem of this two sample size determination methods has not been addressed in the literature. The proposed analysis framework for ordinal data based on the latent variable model can provide convenient test on equal scale assumption and give the estimates of the thresholds. The proposed sample size determination methods based on the latent variable model provides good modification for these cases.

Case	Specification of true parameters			Treatment effect				Sample size			
	$F(x;\mu_1,\sigma_1)$	$F(x;\mu_2,\sigma_2)$	$(\tau_1,\cdots,\tau_{K-1})$	$\hat{\Delta}_{norm}$	$\hat{\Delta}_{logis}$	$\hat{\Delta}_{whd}$	$\hat{\Delta}_{wmw}$	N_{norm}	N_{logis}	N_{whd}	N_{wmw}
	$N(0,1)$	$N(0.15,0.5)$	$(0.5,1.0)$	0.148	0.293	0.462	0.453	3759	2348	720	723
(A)	$N(0,1)$	$N(0.20,0.5)$	$(0.5,1.0)$	0.198	0.358	0.301	0.468	1979	1504	1629	1634
	$N(0,1)$	$N(0.25,0.5)$	$(0.5,1.0)$	0.25	0.427	0.131	0.486	1201	1026	8254	8281
	$N(0,1)$	$N(0.15,2)$	$(0.5,1.0)$	0.158	0.225	-0.642	0.578	7921	10738	311	312
(B)	$N(0,1)$	$N(0.20,2)$	$(0.5,1.0)$	0.198	0.292	-0.676	0.583	4904	6200	279	280
	$N(0,1)$	$N(0.25,2)$	$(0.5,1.0)$	0.246	0.374	-0.718	0.589	3075	3686	246	247
	$L(0,1)$	$L(0.20,0.5)$	$(0.5,1.0)$	0.114	0.192	0.236	0.471	4609	4003	2319	2326
(C)	$L(0,1)$	$L(0.25,0.5)$	$(0.5,1.0)$	0.150	0.246	0.132	0.484	2564	2368	7262	7285
	$L(0,1)$	$L(0.30,0.5)$	$(0.5,1.0)$	0.179	0.291	0.047	0.494	1756	1667	55730	55918
	$L(0,1)$	$L(0.25,2)$	$(0.5,1.0,1.5)$	0.165	0.254	-0.530	0.570	5247	5929	424	426
(D)	$L(0,1)$	$L(0.30,2)$	$(0.5,1.0,1.5)$	0.189	0.294	-0.551	0.573	3938	4371	392	394
	$L(0,1)$	$L(0.40,2)$	$(0.5,1.0,1.5)$	0.254	0.402	-0.603	0.580	2049	2203	325	326
	$L(0,1)$	$L(-0.30,2)$	$(-0.5,0.0,0.5)$	-0.197	-0.316	0.168	0.476	2946	2939	3871	3884
(E)	$L(0,1)$	$L(0.30,2)$	$(-0.5,0.0,0.5)$	0.197	0.316	-0.168	0.524	2945	2939	3871	3884
	$L(0,1)$	$L(0.40,2)$	$(-0.5,0.0,0.5)$	0.248	0.397	-0.212	0.530	1892	1885	2442	2450
	$L(0,2)$	$L(0.40,1.5)$	$(0.5,1.0,1.5)$	0.128	0.204	-0.123	0.517	2863	2842	7713	7739
	$L(0,2)$	$L(0.40,1.5)$	$(0.0,0.5,1.0)$	0.126	0.201	-0.192	0.527	2322	2315	3058	3068
(F)	$L(0,2)$	$L(0.40,1.5)$	$(-0.5,0.0,0.5)$	0.128	0.204	-0.263	0.537	2193	2187	1625	1630
	$L(0,2)$	$L(0.40,1.5)$	$(-1.,-0.5,0.0)$	0.124	0.195	-0.324	0.544	2806	2837	1103	1106
	$L(0,2)$	$L(0.40,1.5)$	$(-1.5,-1.0,-0.5)$	0.123	0.185	-0.394	0.550	4093	4422	800	802
	$L(0,1)$	$L(-0.20,1)$	$(-0.5,0.0,0.5)$	-0.124	-0.199	0.198	0.471	2733	2733	2714	2723
(G)	$L(0,1)$	$L(0.20,1)$	$(-0.5,0.0,0.5)$	0.124	0.199	-0.198	0.529	2733	2733	2714	2723
	$L(0,1)$	$L(0.40,1)$	$(-0.5,0.0,0.5)$	0.251	0.403	-0.399	0.559	700	695	670	672

Table 4.6: The sample size determination of different methods for given ordinal responses

4.7 Conclusion

In this chapter, we proposed a unified framework for the analysis of the latent variable model, where the underlying random variables may have any distributions belonging to the L-S family. On the basis of the proposed two-step estimation procedure, the locations and scales characterizing different treatments can be freely estimated. Subsequently, many statistical inferences can be conveniently conducted based on the proposed method. This framework subsumes the mostly adopted latent variable models in the literature, such as the latent normal model and the latent logistic model. Moreover, this unified framework facilitates the generalization to multiple treatments.

Based on such an analysis framework for the latent variable model, the usual treatment effect measures for comparing the treatments with ordinal responses can be interpreted in a unified manner. Typically, a new interpretation of the proportional odds assumption is given. The corresponding testing methods for the assumption based on such interpretation are also provided. Moreover, based on the latent variable model, the crucial assumption on equal dispersion for the WMW method and the Whd method can be conveniently examined.

Two important latent variable methods for treatment comparison, the *LNorm* method and the *LLogis* method, are detailed illustrated for the comparison of two treatments with ordinal responses. The corresponding sample size determination methods are also proposed. The proposed *LLogis* method is equivalent to the Whitehead (1993) method when the two treatments have identical scales. However, when the treatments have different scales, the existing sample size determination methods will be quite questionable. This problem has not been widely recognized in the literature. Our proposed sample size determination method can accommodate the difference in the scales of different treatments.

Chapter 5

Future Research

In this thesis, we have developed several statistical methods for the comparison of treatments with ordered categorical responses. We mainly addressed three types of treatment comparison issues (TC1, TC2, and TC3) as illustrated in Chapter 1. On the basis of the latent variable models, the proposed methods have easy and straightforward interpretations, and many further statistical inferences can be conveniently conducted. Based upon the present work, here we list several possible topics for future research.

(I) The modeling of longitudinal ordinal responses with more than two repeated measurements

In Chapter 2, we have developed a flexible modeling method for longitudinal ordinal responses with measurements at two time points. This method can be generalized to analyze the longitudinal data with more than two repeated measurements. In this case, not only the linear, but also the non-linear growth curve can be analyzed. Because of the appearance of the high dimensional correlation structure, the computational burden for the likelihood method will increase. In this case, the Bayesian method based on MCMC sampling can serve as an alternative.

(II) The pairwise comparisons with ordinal responses

In Chapter 3, we considered the multiple comparison of several treatments with a control. In multiple comparison, the pairwise comparison is another interesting topic of study, see e.g. Tukey (1953); Hayter (1984, 1989); Cheung and Chan (1996);

Cheung, Wu, and Quek (2003). For treatments with ordered categorical responses, we have derived the variances and the correlation structure of the mean treatment effects in a closed form. So, we may conduct pairwise comparison for treatments with ordinal responses by constructing the corresponding simultaneous confidence intervals.

(III) The covariates-adjusted treatment comparisons

In Chapter 4, we have built a general framework for treatment comparison. The present work can be regarded as the analysis of variance (ANOVA) for ordinal responses. Further extension of the proposed analysis framework for latent variable model is possible, such as the inclusion of covariates in the model. Such covariates-adjusted treatment comparison, which can be regarded as the analysis of covariance (ANCOVA) for ordinal responses, deserves further research. In future studies, some important issues should be addressed, such as variable selection and the specification of the functional form, as illustrated in Tutz (2003), Leon, Tsiatis, and Davidian (2003), and Schacht, Bogaerts, Bluhmki, and Lesaffre (2008).

(IV) The sample size determination in multiple comparisons for ordinal responses

When the responses of the treatments are continuous, the sample size determination for multiple comparison procedures (MCP) has been derived, see, Hayter and Tamhane (1991), Liu (1997), Dunnett, Horn, and Vollandt (2001), and Kwong, Cheung, and Wen (2009). So, it is a natural requirement to consider the sample size determination of the MCPs for ordinal responses.

For the topic of sample size determination, its objective is to determine the smallest total sample size for each MCP to guarantee the specified power requirement for *given* design. The designs of experiments that are usually considered in such study are the balance allocation design and the square-root allocation design. Some other designs of experiments are also possible.

(V) The design of experiments

The *optimal design* of experiments aims to optimize some function of the information obtained in the experiment. For model-based experimental design, a standard

and appropriate measure of information is the Fisher information matrix. For example, Perevozskaya, Rosenberger, and Haines (2003) considered the optimal design for treatments with ordinal responses based on the proportional odds model. Since we have derived the Fisher information matrix in a closed form, the optimal design for treatments comparison with ordinal responses can be considered.

The *response-adaptive design* targets the optimal allocation under multiple objectives, which usually include both the power objective and the ethical objective (assign more patients to better treatments). Since the optimal solution of adaptive design is often a function of unknown parameters, the sequential randomization procedures are usually adopted. Most attention has been focused on binary responses and continuous responses in the literature. For example, Rosenberger *et al.* (2001), Hu and Rosenberger (2003), Tymofyeyev, Rosenberger, and Hu (2007) considered the response-adaptive design for binary response experiments; Zhang and Rosenberger (2006) and Zhu and Hu (2009) considered the response-adaptive design for continuous experiments. Little work has been done on adaptive design for ordinal response. Based on our proposed analysis framework for ordinal responses, the adaptive design for ordinal responses can be exploited.

Remark:

In Section 4.3.1, we proposed the two-step estimation procedure for the general latent variable model. The advantage of this estimation procedure is obvious. Based on this two-step estimation procedure, the location and scale parameters characterizing different treatments can be freely estimated. However, this estimation method may be criticized, since the thresholds are estimated only based on one treatment. A better method is to estimate the thresholds based on the overall likelihood.

Following the suggestion of Professor Anthony Hayter, We propose the following iterative algorithm, where the two-step estimation procedure proposed in Section 4.3.1 will serve as the initial step (Step 0) in the algorithm.

Algorithm 5.1:

- Step 0 (S0):

  S0-1: Determine the values of the thresholds based on treatment s, and let $X_s \sim F(x;0,1)$, $1 \leq s \leq G$.

  S0-2: Obtain the estimates of the location and scale parameters that characterize different treatments, with the thresholds fixed at the values determined in step S0-1.

- Step 1 (S1):

  S1-1: estimate the thresholds based on the *overall likelihood*, with the location and scale parameters fixed at the values obtained in step S0-2.

  S1-2: estimate the location and scale parameters with the thresholds fixed at the values determined in step S1-1.

- Step 2 (S2):

  repeat Step 1 (S1) with the corresponding parameters fixed at the previous latest sub step.

- ...

Our calculation shows that this iterative algorithm converges extremely fast. Usually it achieves convergence in Step 1 (S1), and the Step 2 (S2) will have no effect in the sense that the updated values of the parameters are nearly the same when compared with those obtained by Step 1 (S1). A prominent advantage of this algorithm is that, for different selection of treatment s in Step 0 (S0), this algorithm will produce exactly the same testing results. Moreover, it is not difficult to find that this algorithm actually achieves a full maximum likelihood estimation of all parameters involved in the latent variable model by a conditional maximum likelihood method as illustrated in Meng and Rubbin (1993). Nevertheless, we want to point out that the values of the test statistics that are obtained based on this iterative algorithm are extremely close to those obtained by only Step 0 (S0). So,

for the comparison of only two treatments or the multiple comparison with control (as in Chapter 3), the two-step estimation procedure is sufficient. The advantages of this iterative algorithm may be more prominent when it is adopted to perform more complicated inferences, such as pairwise comparison or the covariates-adjusted treatment comparisons.

The estimation of the thresholds based on the overall likelihood that is mentioned in the algorithm is given as follows. Consider the general latent variable model proposed in Section 4.3.1. The overall log-likelihood function for $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_{K-1})'$ is

$$L(\boldsymbol{\tau}) = \sum_{i=1}^{G}\sum_{k=1}^{K} n_{ik}\log(\pi_{ik}(\boldsymbol{\tau})), \tag{5.1}$$

where

$$\pi_{ik}(\boldsymbol{\tau}) = F_0(\frac{\tau_k - \mu_i}{\sigma_i}) - F_0(\frac{\tau_{k-1} - \mu_i}{\sigma_i}), \tag{5.2}$$

with μ_i and σ_i being fixed parameters. Under regularity conditions, the expected (Fisher) information matrix is given by

$$I(\boldsymbol{\tau}) = E_{\boldsymbol{\tau}}\{\frac{\partial L(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}} \cdot \frac{\partial L(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}'}\}.$$

First note the score function of (5.1) is

$$\frac{\partial L(\boldsymbol{\tau})}{\partial \tau_k} = \sum_{i=1}^{G}\frac{1}{\sigma_i}\left(\frac{n_{ik}}{\pi_{ik}} - \frac{n_{i(k+1)}}{\pi_{i(k+1)}}\right) f_0(\frac{\tau_k - \mu_i}{\sigma_i}), \quad k = 1, \cdots, K-1, \tag{5.3}$$

where $f_0(x)$ is the pdf of the standard distribution function $F_0(x)$.

Then, consider the diagonal elements of $I(\boldsymbol{\tau})$. From the score function,

$$\begin{aligned}
E[\frac{\partial L(\boldsymbol{\tau})}{\partial \tau_k}]^2 &= \sum_{i=1}^{G}\frac{1}{\sigma_i^2}E\left(\frac{n_{ik}}{\pi_{ik}} - \frac{n_{i(k+1)}}{\pi_{i(k+1)}}\right)^2 f_0^2(\frac{\tau_k - \mu_i}{\sigma_i}) \\
&+ \sum_{h\neq l}\frac{1}{\sigma_h}\frac{1}{\sigma_l}E\left(\frac{n_{hk}}{\pi_{hk}} - \frac{n_{h(k+1)}}{\pi_{h(k+1)}}\right)E\left(\frac{n_{lk}}{\pi_{lk}} - \frac{n_{l(k+1)}}{\pi_{l(k+1)}}\right) f_0(\frac{\tau_k - \mu_h}{\sigma_h}) f_0(\frac{\tau_k - \mu_l}{\sigma_l}) \\
&= \sum_{i=1}^{G}\frac{1}{\sigma_i^2}\left(\frac{E(n_{ik}^2)}{\pi_{ik}^2} + \frac{E(n_{i(k+1)}^2)}{\pi_{i(k+1)}^2} - 2\frac{E(n_{ik}n_{i(k+1)})}{\pi_{ik}\pi_{i(k+1)}}\right) f_0^2(\frac{\tau_k - \mu_i}{\sigma_i}) + 0.
\end{aligned}$$

From the property of multinomial distribution, the above equation

$$= \sum_{i=1}^{G} \frac{1}{\sigma_i^2} \left(\frac{n_i \pi_{ik}(1-\pi_{ik}) + n_i^2 \pi_{ik}^2}{\pi_{ik}^2} + \frac{n_i \pi_{i(k+1)}(1-\pi_{i(k+1)}) + n_i^2 \pi_{i(k+1)}^2}{\pi_{i(k+1)}^2} - 2(n_i^2 - n_i) \right)$$
$$\cdot f_0^2\left(\frac{\tau_k - \mu_i}{\sigma_i}\right)$$
$$= \sum_{i=1}^{G} \frac{n_i}{\sigma_i^2} \left(\frac{1}{\pi_{ik}} + \frac{1}{\pi_{i(k+1)}} \right) f_0^2\left(\frac{\tau_k - \mu_i}{\sigma_i}\right). \tag{5.4}$$

Now, we consider the off-diagonal elements of $I(\boldsymbol{\tau})$, for $1 \le h \ne l \le K-1$,

$$E\left[\frac{\partial L(\boldsymbol{\tau})}{\partial \tau_h} \frac{\partial L(\boldsymbol{\tau})}{\partial \tau_l}\right] = E\left[\sum_{i=1}^{G} \frac{1}{\sigma_i} \left(\frac{n_{ih}}{\pi_{ih}} - \frac{n_{i(h+1)}}{\pi_{i(h+1)}} \right) f_0\left(\frac{\tau_h - \mu_i}{\sigma_i}\right) \right]$$
$$\cdot \left[\sum_{i=1}^{G} \frac{1}{\sigma_i} \left(\frac{n_{il}}{\pi_{il}} - \frac{n_{i(l+1)}}{\pi_{i(l+1)}} \right) f_0\left(\frac{\tau_l - \mu_i}{\sigma_i}\right) \right]$$
$$= \sum_{i=1}^{G} \frac{1}{\sigma_i^2} E\left(\frac{n_{ih}}{\pi_{ih}} - \frac{n_{i(h+1)}}{\pi_{i(h+1)}} \right) \left(\frac{n_{il}}{\pi_{il}} - \frac{n_{i(l+1)}}{\pi_{i(l+1)}} \right) f_0\left(\frac{\tau_h - \mu_i}{\sigma_i}\right) f_0\left(\frac{\tau_l - \mu_i}{\sigma_i}\right)$$
$$= \sum_{i=1}^{G} \frac{1}{\sigma_i^2} E\left(\frac{n_{ih} n_{il}}{\pi_{ih}\pi_{il}} - \frac{n_{ih} n_{i(l+1)}}{\pi_{ih}\pi_{i(l+1)}} - \frac{n_{i(h+1)} n_{il}}{\pi_{i(h+1)}\pi_{il}} + \frac{n_{i(h+1)} n_{i(l+1)}}{\pi_{i(h+1)}\pi_{i(l+1)}} \right) f_0\left(\frac{\tau_h - \mu_i}{\sigma_i}\right) f_0\left(\frac{\tau_l - \mu_i}{\sigma_i}\right)$$

Note that the expectation involved in the above equation will equal to 0 if $|h-l| > 1$. For the case $h = l+1$,

$$E\left[\frac{\partial L(\boldsymbol{\tau})}{\partial \tau_h} \frac{\partial L(\boldsymbol{\tau})}{\partial \tau_l}\right]$$
$$= \sum_{i=1}^{G} \frac{1}{\sigma_i^2} \left(n_i^2 - n_i - \frac{E(n_{ih}^2)}{\pi_{ih}^2} \right) f_0\left(\frac{\tau_h - \mu_i}{\sigma_i}\right) f_0\left(\frac{\tau_{h-1} - \mu_i}{\sigma_i}\right)$$
$$= \sum_{i=1}^{G} \frac{1}{\sigma_i^2} \left(n_i^2 - n_i - \frac{n_i \pi_{ih}(1-\pi_{ih}) + n_i^2 \pi_{ih}^2}{\pi_{ih}^2} \right) f_0\left(\frac{\tau_h - \mu_i}{\sigma_i}\right) f_0\left(\frac{\tau_{h-1} - \mu_i}{\sigma_i}\right)$$
$$= -\sum_{i=1}^{G} \frac{n_i}{\sigma_i^2} \frac{1}{\pi_{ih}} f_0\left(\frac{\tau_h - \mu_i}{\sigma_i}\right) f_0\left(\frac{\tau_{h-1} - \mu_i}{\sigma_i}\right), \quad h = 2, \cdots, K-1. \tag{5.5}$$

For the case $h = l-1$, a similar expression for $E[\frac{\partial L(\boldsymbol{\tau})}{\partial \tau_h} \frac{\partial L(\boldsymbol{\tau})}{\partial \tau_l}]$ can be derived. Thus, we derive the form of the Fisher information matrix $I(\boldsymbol{\tau})$. Numerical method can be applied to obtain the MLE of $\boldsymbol{\tau}$ and the corresponding covariance matrix.

Bibliography

[1] Agresti, A. (1983). A survey of strategies for modeling cross classification having ordinal variables. *Journal of the American Statistical Association* **78**, 184-198.

[2] Agresti, A. (1984) *Analysis of Ordinal Categorical Data.* Wiley: New York.

[3] Agresti, A. (1989). A survey of models for repeated ordered categorical response data *Statistics in Medicine* **8**, 1209-1224.

[4] Agresti, A. (1999). Modelling ordered categorical data: Recent advances and future challenges. *Statistics in Medicine* **18**, 2191-2207.

[5] Agresti, A. (2002). *Categorical Data Analysis (Second Edition).* Wiley: New York.

[6] Anderson, J.A. (1984). Regression and Ordered categorical variables (with discussion). *Journal of the Royal Statistical Society, Series B*, **46**, 1-30

[7] Anderson, J.A. and Philips, P.R. (1981). Regression, discrimination and measurement models for ordered categorical variables. *Applied Statistics* **30**, 22-31.

[8] Bartholomew, D.J. (1980). Discussion on Regression models for ordinal data (by P. McCullagh). *Journal of the Royal Statistical Society, Series B*, **42**, 127-129.

[9] Bechhofer, R.E. and Dunnett, C.W. (1988). Tables of percentage points of multivariate t distributions. In *Selected Tables in Mathematical Statistics* **11**. Providence, Rhode Island. American Mathematical Society.

[10] Bekele, B.N. and Thall, P.F. (2004). Dose-finding based on multiple toxicities in a soft tissue scarcoma trial. *Journal of the American Statistical Association* **99**, 26-35.

[11] Bender, T. and Grouven, U. (1998). Using binary logistic regression models for ordinal data with non-proportional odds. *Journal of Clinical Epidemiology*, **51**, 809-816.

[12] Bollen, K.A. and Curran, P.J. (2006). *Latent Curve Models: A Structural Equation Perspective*. Wiley: New York.

[13] Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, **46**, 1171-1178.

[14] Chan, K.S. and Munoz-Hernandez, B. (2003) A generalized linear model for repeated ordered categorical response data. *Statistica Sinica* **13**, 207-226.

[15] Cheung, S.H. and Chan, W.S. (1996). Simultaneous confidence intervals for pairwise multiple comparisons in a two-way unbalanced design. *Biometrics* **52**, 463-472.

[16] Cheung, S.H., Wu, K.H., and Quek, A.L. (2003). Pairwise comparisons in each of several groups with heterogenous group variances. *Biometrical Journal* **45**, 325-334.

[17] Cox, D.R. (1970). *The Analysis of Binary Data*. Chapman & Hall: London.

[18] Dark, R., Bolland, K., and Whitehead, J. (2003). Statistical methods for ordered categorical data based on a constrained odds model. *Biometrical Journal* **45**, 453-470.

[19] Diem, S., Grady, D., Quan, J., Vittinghoff, E., Wallace, R., Hanes, V., and Ensrud, K. (2006). Effects of ultralow-dose transdermal estradiol on postmenopausal symptoms in women aged 60 to 80 years *Menopause* **13**, 130-138.

[20] Duncan, T.E., Duncan, S.C., Strycker, L.A., Li, F., and Alpert, A. (1999) *An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues, and Applications*. Lawrence Erlbaum Associates: London.

[21] Dunnett, C.W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* **50**, 1096-1121.

[22] Dunnett, C W (1989). Multivariate normal probability integrals with product correlation structure. *Applied Statistics* **38**, 564-579.

[23] Dunnett, C.W. and Gent, M. (1977). Significance testing to establish equivalence between treatments with special reference to data in the form of 2×2 tables. *Biometrics* **33**, 593-602.

[24] Dunnett, C.W., Horn, M., and Vollandt, R. (2001). Sample size determination in step-down and step-up multiple tests for comparing treatments with a control. *Journal of Statistical Planning and Inference* **97**, 367-384.

[25] Dunnett, C W. and Tamhane, A.C. (1991). Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. *Statistics in Medicine* **10**, 939-947.

[26] Dunnett, C.W. and Tamhane, A.C. (1992). A step-up multiple test procedure. *Journal of the American Statistical Association* **87**, 162-170.

[27] Fujii, Y. and Itakura, M. (2009). A comparison of pretreatment with fentanyl and lidocaine preceded by venous occlusion for reducing pain on injection of propofol: A prospective, randomized, double-blind, placebo-controlled study in adult Japanese surgical patients. *Clinical Therapeutics* **31**, 2107-2112

[28] Hayter, A.J. (1984). A proof of the conjecture that the Tukey-kramer multiple comparisons procedure is conservative. *The Annals of Statistics* **12**, 61-75.

[29] Hayter, A.J. (1989). Pairwise comparison of generally correlated means. *Journal of the American Statistical Association* **84**, 208-213.

[30] Hayter, A.J. and Tamhane, A.C. (1991). Sample size determination for step-down multiple test procedures: orthogonal contrasts and comparisons with a control *Journal of Statistical Planning and Inference* **27**, 271-290.

[31] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800-802.

[32] Hochberg, Y and Tamhane, A. (1987). *Multiple Comparison procedures.* New York· Wiley.

[33] Holm, S (1979) A simple sequentially rejective multiple test procedure *Scandinavian Journal of Statistics* **6**, 65-70

[34] Horn, M and Dunnett, C W (2004) Power and sample size determination of stepwise FWE and FDR controlling test procedures in the normal many-one case In *Recent Developments in Multiple Comparison Procedures*, Y Benjamini, F Bretz, and S Sarkar (eds), 48-64 Beachwood, Ohio, USA Institute of Mathematical Statistics

[35] Hsu, J C (1996) *Multiple Comparisons Theory and Methods* New York Chapman and Hall

[36] Hu, F and Rosenberger, W F (2003) Optimality, variability, power Evaluating response-adaptive randomization procedures for treatment comparisons *Journal of the American Statistical Association* **98**, 671-678

[37] Joreskog, K G and Sorbom, D (1999) *PRELIS 2 Users Reference Guide* Chicago Scientific Software International

[38] Joreskog, K G and Sorbom, D (2004) *LISREL 8* Chicago Scientific Software International URL `http //www ssicentral com/lisrel/mainlis htm`

[39] Kalbfleisch, J D and Lawless, J F (1985) The analysis of panel data under a Markov assumption *Journal of the American Statistical Association* **80**, 862-871

[40] Koo, S W , Cho, S J , Kim, Y K , Ham, K D , and Hwang, J H (2006) Small-dose ketamine reduces the pain of propofol injection *Anesthesia & Analgesia* **103**, 1444-1447

[41] Kwong, K S , Cheung, S H , and Wen, M J (2010) Sample size determination in step-up testing procedures for multiple comparisons with a control *Statistics in medicine* **29**, 2743-2756

[42] Kwong, K S and Liu, W (2000) Calculation of critical values for Dunnett and Tamhane's step-up multiple test procedure *Statistics & Probability Letters* **49**, 411-416

[43] Lee, S Y (2007) *Structural Equation Modelling A Bayesian Approach* Chichester, England Wiley

[44] Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* California: Holden-Day.

[45] Leon, S., Tsiatis, A.A., and Davidian, M. (2003). Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics* **59**, 1046-1055.

[46] Leon-Novelo, L.G., Zhou, X., Bekele, B.N., and Müller, P. (2010). Assessing toxicities in a clinical trial: Bayesian inference for ordinal data nested within categories. *Biometrics* **66**, 966-974.

[47] Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

[48] Liu, J.P., Hsueh, H.M., Hsieh, E., and Chen, J.J. (2002). Tests for equivalence or non-inferiority for paired binary data. *Statistics in Medicine* **21**, 231-245.

[49] Liu, W. (1997). On sample size determination of Dunnett's procedure for comparing several treatments with a control. *Journal of Statistical Planning and Inference* **62**, 255-261.

[50] Lu, T.Y., Poon, W.Y., and Tsang, Y.F. (2011). Latent growth curve modeling for longitudinal ordinal responses with applications. *Computational Statistics and Data Analysis* **55**, 1488-1497.

[51] Lu, T.Y., Poon, W.Y., and Cheung, S.H. (2011). Multiple testing of several treatments with a control for ordered categorical responses. Submitted.

[52] Lui, K.J., Cumberland, W.G., 2001. A test procedure of equivalence in ordinal data with matched-pairs. *Biometrical Journal* **43**, 977-983.

[53] Lui, K.J. and Zhou, X.H. (2004). Testing non-inferiority (and equivalence) between two diagnostic procedures in paired-sample ordinal data. *Statistics in Medicine* **23**, 545-559.

[54] Mann, H.B. and Whitney, D.R (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* **18**, 50-60.

[55] Marcus, R., Peritz, E., and Gabriel, K.R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655-660.

[56] Meng, X.L. and Rubbin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80**, 267-278.

[57] McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of Royal Statistical Society B* **42**, 109-142.

[58] Nam, J.M. (1997). Establishing equivalence of two treatments and sample size requirements in matched-pairs design. *Biometrics* **53**, 1442-1430.

[59] Neale, M.C., Boker, S.M., Xie, G., and Maes, H.H. (1999). *Mx: Statistical Modeling.* Box 126 MCV, Richmond, VA23298: Department of Psychiatry, 5th Edition. `http://www.vcu.edu/mx/`.

[60] Pédrono, G., Thiébaut, R., Alioum, A., Lesprit, P., Fritzell, B., Lèvy, Y., Chêne, G. (2009). A new endpoint definition improved clinical relevance and statistical power in a vaccine trial. *Journal of Clinical Epidemiology* **62**, 1054-1061.

[61] Perevozskaya, I., Rosenberger, W.F., and Haines, L.M. (2003). Optimal design for the proportional odds model. *The Canadian Journal of Statistics* **31**, 225-235.

[62] Peterson, B. and Harrell, F.E. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics* **39**, 205-217.

[63] Poon, W.Y. (2004). A latent normal distribution model for analysing ordinal responses with applications in meta-analysis. *Statistics in Medicine* **23**, 2155-2172.

[64] Pratt, J.W. (1964). Robustness of some procedures for the two-sample location problem. *Journal of the American Statistical Association*, **59**, 665-680.

[65] Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033-1048.

[66] Qu, Y., Piedmonte, M.R., and Medendorp, S.V. (1995). Latent variable modes for clustered ordinal data. *Biometrics* **51**, 268-275.

[67] Rabbee, N., Coull, B.A., and Mehta, C. (2003). Power and sample size for ordered categorical data. *Statistical Methods in Medical Research*, **12**, 73-84.

[68] Rom, D.M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika* **77**, 663-665.

[69] Rosenberger, W.F., Stallard, N., Ivanova, A., Harper, C.N., and Ricks, M.L. (2001). Optimal adaptive designs for binary response trials. *Biometrics* **57**, 909-913.

[70] Ryu, E. and Agresti, A. (2008). Modeling and inference for an ordinal effect size measure. *Statistics in Medicine*, **27**, 1703-1717.

[71] Schacht, A., Bogaerts, K., Bluhmki, E., and Lesaffre, E. (2008). A new nonparametric approach for baseline covariates adjustment for two-group comparative studies. *Biometrics* **64**, 1110-1116.

[72] Tang, M.L. and Poon, W.Y. (2007). Statistical inference for equivalence trials with ordinal responses: A latent normal distribution approach. *Computational Statistics and Data Analysis* **51**, 5918-5926.

[73] Todem, D., Kim, K.M., and Lesaffre, E. (2007). Latent-variable models for longitudinal data with bivariate ordinal outcomes. *Statistics in Medicine* **26**, 1034-1054.

[74] Tukey, J.W. (1953). *The Problem of Multiple Comparisons.* Unpublished report, Princeton University, Princeton, New Jersey.

[75] Tutz, G. (2003). Generalized semiparametrically structured ordinal models. *Biometrics* **59**, 263-273.

[76] Tymofyeyev, Y., Rosenberger, W.F., and Hu, F. (2007). Implementing optimal allocation in sequential binary response experiments. *Journal of the American Statistical Association* **102**, 224-234.

[77] Wang, W.W.B , Mehrotra, D.V , Chan, I S.F., and Heyse, J.F. (2006) Statistical considerations for non-inferiority/equivalence trials in vaccine development. *Journal of Biopharmaceutical Statistics* **16**, 429-441.

[78] Wetherill, G.B. (1960). The Wilcoxon test and non-null hypotheses. *Journal of the Royal Statistical Society, Series B*, **22**, 402-418.

[79] Whitehead, J. (1993). Sample size calculation for ordered categorical data. *Statistics in Medicine* **12**, 2257-2271.

[80] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics* **1**, 80-83.

[81] Zeger, S.L. and Liang, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121-130.

[82] Zhao, Y.D., Rahardja, D., and Qu, Y. (2008). Sample size calculation for the Wilcoxon-Mann-Whitney test adjusting for ties. *Statistics in Medicine*, **27**, 462-468.

[83] Zhang, L. and Rosenberger, W.F. (2006). Response-adaptive randomization for clinical trials with continous outcomes. *Biometrics* **62**, 562-569.

[84] Zhu, H. and Hu, F. (2009). Implementing optimal allocation for sequential continuous responses with multiple treatments. *Journal of Statistical Planning and Inference* **139**, 2420-2430.

Appendix A

The Mx Input Script for LCM

Appendix A-1:

```
!Sample Mx input program for LCM
!For data set of active drug and placebo in Table 2.1
!To get the results in Table 2.3 and the L0 test
#define nvar=2        ! No of variables
#define nthd=3        ! No of thresholds
#define nlat=2        ! No of latent variables (L.Vs)
#define n1=119        ! No of observations in active drug group
#define n2=120        ! No of observations in placebo group
#NGroups 3
Group 1: parameters
Calculation
Begin Matrices;
T Full nvar nthd      ! Threshold matrix, invariant across groups
L Full nvar nlat      ! Factor loading matrix
K Full nlat 1         ! Mean vector of L.Vs for drug group
E Symm nvar nvar Fix  ! Error variance matrix for drug group
F Symm nlat nlat      ! Covariance matrix of L.Vs for drug group
G Full nlat 1         ! Mean vector of L.Vs for placebo group
D Symm nvar nvar Fix  ! Error variance matrix for placebo group
```

```
H Symm nlat nlat      ! Covariance matrix of L.V's for placebo group
O Full 1 nthd Fix     ! The operation matrix
End Matrices;
Specify T
1 2 3
1 2 3
Specify L
0 0
0 0
Specify K
0
4
Specify F
0
5
Specify G
11
7
Specify H
12
8 9
Matrix T
-1 0 1
-1 0 1
Matrix L
1 0
1 1
Matrix K
0
```

```
0
Matrix E
0
0 0
Matrix F
1
0 1
Matrix G
0
0
Matrix D
0
0 0
Matrix H
1
0 1
Matrix O
1 1 1
End
Group 2: Active drug
Data NI=nvar NO=n1
CTable 4 4
7    4     1     0
11   5     2     2
13   23    3     1
9    17    13    8
Matrices=Group 1
Threshold T-(L*K*O);
Covariance L*F*L'+E;
```

```
Option RS
Option SErrors
End
Group 3: Placebo group
Data NI=nvar NO=n2
CTable 4 4
7    4     2     1
14   5     1     0
6    9     18    2
4    11    14    22
Matrices=Group 1
Threshold T-(L*G*O);
Covariance L*H*l'+D;
Option RS
Option SErrors
End
```

Appendix A-2:

```
!Sample Mx program for LCM with covariates
!For data set of active drug and placebo in Table 2.1
#define nvar=3        ! No of variables
#define nthd=3        ! No of thresholds
#define nlat=2        ! No of latent variables
#define n=239         ! No of observations
Ordinal data analysis
DAta NI=nvar NO=n NG=1
ORdinal_data file=rawdata.txt
Begin Matrices;
T Full nthd nvar  Free
```

```
L Full 2 2
X Full 1 1  Fix      ! mean of x
S Full 1 1  Fix      ! Variance of x
N Full nlat 1        ! Gamma vector
K Full nlat 1        ! The mean vector of L.Vs
E Symm 2  2  Fix     ! The error variance matrix
F Symm nlat nlat     ! The covariance matrix of L.Vs
O Full nthd 1  Fix   ! The operation matrix
End Matrices;
Specify T
1 1 4
2 2 4
3 3 4
Specify  L
0 0
0 0
Specify K
0
5
Specify F
0
6  7
Specify N
10
11
Matrix T
-1  -1    .5
0    0    .5
1    1    .5
```

```
Matrix L
1  0
1  1
Matrix K
0
0
Matrix F
1
0 1
Matrix 0
1
1
1
Matrix X 0.4979079   ! Fixed at the estimate of the mean of x
Matrix S 0.251046    ! Fixed at the estimate of the variance of x
Matrix N
0
0
Begin Algebra;
M=(L*(K+N*X)_ X )'@0;
C=L*(F+N*S*N')*L'+E | L*N*S_
          S*N'*L'  |   S ;
End Algebra;
Threshold T-M;
Covariance C;
Option func=1.E-10
Option SE
End
```

Appendix B

The Proof of Theorem 3.2 and Theorem 3.3

Proof of Theorem 3.2

The log-likelihood function for step 2 of the two-step estimation procedure can be written as

$$L(\boldsymbol{\theta}_0) = \sum_{i=0}^{G}\sum_{k=1}^{K} n_{ik}\log(\pi_{ik}(\boldsymbol{\theta}_0)), \tag{B.1}$$

where

$$\pi_{ik}(\boldsymbol{\theta}_0) = \Phi(\frac{\tau_k - \mu_i}{\sigma_i}) - \Phi(\frac{\tau_{k-1} - \mu_i}{\sigma_i}), \tag{B.2}$$

and $\boldsymbol{\theta}_0 = (\mu_0, \cdots, \mu_G, \sigma_0, \cdots, \sigma_G)'$ is the vector containing all unknown parameters. Under regularity conditions, the expected (Fisher) information matrix is given by

$$I(\boldsymbol{\theta}_0) = E_{\boldsymbol{\theta}_0}\{\frac{\partial L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\cdot\frac{\partial L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'}\} = \begin{bmatrix} A & B \\ B' & C \end{bmatrix},$$

where A, B, and C are $(G+1)\times(G+1)$ matrices with expressions given as follows.

(a) Matrix A:

First consider the diagonal elements of A. From (B.1),

$$[\frac{\partial L(\boldsymbol{\theta}_0)}{\partial \mu_i}]^2 = [\sum_{k=1}^{K}\frac{n_{ik}}{\pi_{ik}}\frac{\partial \pi_{ik}}{\partial \mu_i}]^2 = \sum_{k=1}^{K}(\frac{n_{ik}}{\pi_{ik}}\frac{\partial \pi_{ik}}{\partial \mu_i})^2 + \sum_{h\neq l}\frac{n_{ih}}{\pi_{ih}}\frac{n_{il}}{\pi_{il}}\frac{\partial \pi_{ih}}{\partial \mu_i}\frac{\partial \pi_{il}}{\partial \mu_i}.$$

Then, by equation (B.2) and the properties of the multinomial distribution, we have

$$
\begin{aligned}
& E[\frac{\partial L(\boldsymbol{\theta}_0)}{\partial \mu_i}]^2 \\
= & \sum_{k=1}^{K} \frac{E(n_{ik})^2}{(\pi_{ik})^2} (\frac{\partial \pi_{ik}}{\partial \mu_i})^2 + \sum_{h \neq l} \frac{E(n_{ih} n_{il})}{\pi_{ih} \pi_{il}} \frac{\partial \pi_{ih}}{\partial \mu_i} \frac{\partial \pi_{il}}{\partial \mu_i} \\
= & \sum_{k=1}^{K} \frac{(n_i \pi_{ik}(1-\pi_{ik}) + n_i^2 \pi_{ik}^2)}{(\pi_{ik})^2} (\frac{\partial \pi_{ik}}{\partial \mu_i})^2 + \sum_{h \neq l} \frac{(-n_i \pi_{ih} \pi_{il} + n_i^2 \pi_{ih} \pi_{il})}{\pi_{ih} \pi_{il}} \frac{\partial \pi_{ih}}{\partial \mu_i} \frac{\partial \pi_{il}}{\partial \mu_i} \\
= & \frac{n_i}{\sigma_i^2} \{ \sum_{k=1}^{K} (\frac{1}{\pi_{ik}} + n_i - 1) [\phi(\frac{\tau_k - \mu_i}{\sigma_i}) - \phi(\frac{\tau_{k-1} - \mu_i}{\sigma_i})]^2 \\
& + \sum_{h \neq l} (n_i - 1) [\phi(\frac{\tau_h - \mu_i}{\sigma_i}) - \phi(\frac{\tau_{h-1} - \mu_i}{\sigma_i})] [\phi(\frac{\tau_l - \mu_i}{\sigma_i}) - \phi(\frac{\tau_{l-1} - \mu_i}{\sigma_i})] \} \\
= & \frac{n_i}{\sigma_i^2} \{ \sum_{k=1}^{K} \frac{1}{\pi_{ik}} [\phi(\frac{\tau_k - \mu_i}{\sigma_i}) - \phi(\frac{\tau_{k-1} - \mu_i}{\sigma_i})]^2 \\
& + (n_i - 1) [\sum_{k=1}^{K} \phi(\frac{\tau_k - \mu_i}{\sigma_i}) - \phi(\frac{\tau_{k-1} - \mu_i}{\sigma_i})]^2 \} \\
= & \frac{n_i}{\sigma_i^2} \{ \sum_{k=1}^{K} \frac{1}{\pi_{ik}} [\phi(\frac{\tau_k - \mu_i}{\sigma_i}) - \phi(\frac{\tau_{k-1} - \mu_i}{\sigma_i})]^2 \} \\
= & \frac{n_i}{\sigma_i^2} \cdot \frac{1}{\delta_0(\mu_i, \sigma_i)}.
\end{aligned}
\tag{B.3}
$$

Now, consider the off-diagonal elements of A. For $i \neq j$,

$$
\begin{aligned}
& E[\frac{\partial L(\boldsymbol{\theta}_0)}{\partial \mu_i} \cdot \frac{\partial L(\boldsymbol{\theta}_0)}{\partial \mu_j}] \\
= & E\{ [\sum_{k=1}^{K} \frac{n_{ik}}{\pi_{ik}} \frac{\partial \pi_{ik}}{\partial \mu_i}] [\sum_{k=1}^{K} \frac{n_{jk}}{\pi_{jk}} \frac{\partial \pi_{jk}}{\partial \mu_j}] \} \\
= & \frac{n_i n_j}{\sigma_i \sigma_j} \{ [\sum_{k=1}^{K} (\phi(\frac{\tau_k - \mu_i}{\sigma_i}) - \phi(\frac{\tau_{k-1} - \mu_i}{\sigma_i}))] [\sum_{k=1}^{K} (\phi(\frac{\tau_k - \mu_j}{\sigma_j}) - \phi(\frac{\tau_{k-1} - \mu_j}{\sigma_j}))] \} \\
= & \frac{n_i n_j}{\sigma_i \sigma_j} [\phi(\frac{\tau_K - \mu_j}{\sigma_j}) - \phi(\frac{\tau_0 - \mu_j}{\sigma_j})] [\phi(\frac{\tau_K - \mu_j}{\sigma_j}) - \phi(\frac{\tau_0 - \mu_j}{\sigma_j})] \\
= & 0.
\end{aligned}
$$

The above equation equals 0, because $\tau_0 = -\infty$ and $\tau_K = \infty$ by definition.

(b) Matrix B:

First, consider the diagonal elements of B. We have

$$
\begin{aligned}
& E[\frac{\partial L(\boldsymbol{\theta}_0)}{\partial \mu_i} \cdot \frac{\partial L(\boldsymbol{\theta}_0)}{\partial \sigma_i}] \\
= & \sum_{k=1}^{K} \frac{E(n_{ik})^2}{(\pi_{ik})^2} \frac{\partial \pi_{ik}}{\partial \mu_i} \frac{\partial \pi_{ik}}{\partial \sigma_i} + \sum_{h \neq l} \frac{E(n_{ih} n_{il})}{\pi_{ih} \pi_{il}} \frac{\partial \pi_{ih}}{\partial \mu_i} \frac{\partial \pi_{il}}{\partial \sigma_i} \\
= & \sum_{k=1}^{K} \frac{(n_i \pi_{ik}(1-\pi_{ik}) + n_i^2 \pi_{ik}^2)}{(\pi_{ik})^2} \frac{\partial \pi_{ik}}{\partial \mu_i} \frac{\partial \pi_{ik}}{\partial \sigma_i} + \sum_{h \neq l} \frac{(-n_i \pi_{ih} \pi_{il} + n_i^2 \pi_{ih} \pi_{il})}{\pi_{ih} \pi_{il}} \frac{\partial \pi_{ih}}{\partial \mu_i} \frac{\partial \pi_{il}}{\partial \sigma_i} \\
= & n_i \{ \sum_{k=1}^{K} \frac{1}{\pi_{ik}} \frac{\partial \pi_{ik}}{\partial \mu_i} \frac{\partial \pi_{ik}}{\partial \sigma_i} + (n_i - 1) [\sum_{k=1}^{K} \frac{\partial \pi_{ik}}{\partial \mu_i}][\sum_{k=1}^{K} \frac{\partial \pi_{ik}}{\partial \sigma_i}] \} \\
= & \frac{n_i}{\sigma_i^3} \{ \sum_{k=1}^{K} \frac{1}{\pi_{ik}} \cdot [\phi(\frac{\tau_k - \mu_i}{\sigma_i}) - \phi(\frac{\tau_{k-1} - \mu_i}{\sigma_i})] \\
& \cdot [(\tau_k - \mu_i)\phi(\frac{\tau_k - \mu_i}{\sigma_i}) - (\tau_{k-1} - \mu_i)\phi(\frac{\tau_{k-1} - \mu_i}{\sigma_i})] \} \\
= & \frac{n_i}{\sigma_i^3} \cdot \delta_1(\mu_i, \sigma_i). \qquad \text{(B.4)}
\end{aligned}
$$

For the off-diagonal elements of B, we have

$$
\begin{aligned}
& E[\frac{\partial L(\boldsymbol{\theta}_0)}{\partial \mu_i} \cdot \frac{\partial L(\boldsymbol{\theta}_0)}{\partial \sigma_j}] \\
= & E\{ [\sum_{k=1}^{K} \frac{n_{ik}}{\pi_{ik}} \frac{\partial \pi_{ik}}{\partial \mu_i}][\sum_{k=1}^{K} \frac{n_{jk}}{\pi_{jk}} \frac{\partial \pi_{jk}}{\partial \sigma_j}] \} \\
= & -\frac{n_i n_j}{\sigma_i \sigma_j^2} \{ [\sum_{k=1}^{K} (\phi(\frac{\tau_k - \mu_i}{\sigma_i}) - \phi(\frac{\tau_{k-1} - \mu_i}{\sigma_i}))] \\
& \cdot [\sum_{k=1}^{K} ((\tau_k - \mu_j)\phi(\frac{\tau_k - \mu_j}{\sigma_j}) - (\tau_{k-1} - \mu_j)\phi(\frac{\tau_{k-1} - \mu_j}{\sigma_j}))] \} \\
= & -\frac{n_i n_j}{\sigma_i \sigma_j^2} [\phi(\frac{\tau_K - \mu_i}{\sigma_i}) - \phi(\frac{\tau_0 - \mu_i}{\sigma_i})][(\tau_K - \mu_j)\phi(\frac{\tau_K - \mu_j}{\sigma_j}) - (\tau_0 - \mu_j)\phi(\frac{\tau_0 - \mu_j}{\sigma_j})] \\
= & 0.
\end{aligned}
$$

(c) Matrix C:

The ith diagonal element of C is

$$\begin{aligned}
& E[\frac{\partial L(\boldsymbol{\theta}_0)}{\partial \sigma_i}]^2 \\
= & \sum_{k=1}^{K} \frac{E(n_{ik})^2}{(\pi_{ik})^2}(\frac{\partial \pi_{ik}}{\partial \sigma_i})^2 + \sum_{h \neq l} \frac{E(n_{ih}n_{il})}{\pi_{ih}\pi_{il}} \frac{\partial \pi_{ih}}{\partial \sigma_i} \frac{\partial \pi_{il}}{\partial \sigma_i} \\
= & \frac{n_i}{\sigma_i^4} \cdot \{\sum_{k=1}^{K} \frac{1}{\pi_{ik}} \cdot [(\tau_k - \mu_i)\phi(\frac{\tau_k - \mu_i}{\sigma_i}) - (\tau_{k-1} - \mu_i)\phi(\frac{\tau_{k-1} - \mu_i}{\sigma_i})]^2\} \\
= & \frac{n_i}{\sigma_i^4} \cdot \delta_2(\mu_i, \sigma_i). \qquad \text{(B.5)}
\end{aligned}$$

The off-diagonal elements of C are

$$\begin{aligned}
& E[\frac{\partial L(\boldsymbol{\theta}_0)}{\partial \sigma_i} \cdot \frac{\partial L(\boldsymbol{\theta}_0)}{\partial \sigma_j}] \\
= & E\{[\sum_{k=1}^{K} \frac{n_{ik}}{\pi_{ik}} \frac{\partial \pi_{ik}}{\partial \sigma_i}][\sum_{k=1}^{K} \frac{n_{jk}}{\pi_{jk}} \frac{\partial \pi_{jk}}{\partial \sigma_j}]\} \\
= & \frac{n_i n_j}{\sigma_i^2 \sigma_j^2}\{[\sum_{k=1}^{K}((\tau_k - \mu_i)\phi(\frac{\tau_k - \mu_i}{\sigma_i}) - (\tau_{k-1} - \mu_i)\phi(\frac{\tau_{k-1} - \mu_i}{\sigma_i}))] \\
& \cdot [\sum_{k=1}^{K}((\tau_k - \mu_j)\phi(\frac{\tau_k - \mu_j}{\sigma_j}) - (\tau_{k-1} - \mu_j)\phi(\frac{\tau_{k-1} - \mu_j}{\sigma_j}))]\} \\
= & \frac{n_i n_j}{\sigma_i^2 \sigma_j^2}[(\tau_K - \mu_i)\phi(\frac{\tau_K - \mu_i}{\sigma_i}) - (\tau_0 - \mu_i)\phi(\frac{\tau_0 - \mu_i}{\sigma_i})] \\
& \cdot [(\tau_K - \mu_j)\phi(\frac{\tau_K - \mu_j}{\sigma_j}) - (\tau_0 - \mu_j)\phi(\frac{\tau_0 - \mu_j}{\sigma_j})] \\
= & 0
\end{aligned}$$

Thus, the matrices A, B, and C can be expressed as:

$$\begin{aligned}
A &= \text{diag}\{\frac{n_i}{\sigma_i^2} \cdot \frac{1}{\delta_0(\mu_i, \sigma_i)}, \quad i = 0, \cdots, G\}, \\
B &= \text{diag}\{\frac{n_i}{\sigma_i^3} \cdot \delta_1(\mu_i, \delta_i), \quad i = 0, \cdots, G\}, \\
C &= \text{diag}\{\frac{n_i}{\sigma_i^4} \cdot \delta_2(\mu_i, \sigma_i), \quad i = 0, \cdots, G\}.
\end{aligned}$$

From the results of block matrix decomposition, the inverse of the expected information matrix can be expressed as

$$I^{-1}(\boldsymbol{\theta}_0) = \begin{bmatrix} A^{-1} + A^{-1}BS_A^{-1}BA^{-1} & -A^{-1}BS_A^{-1} \\ -S_A^{-1}BA^{-1} & S_A^{-1} \end{bmatrix},$$

where $S_A = (C - BA^{-1}B)$. As all submatrices of $I(\boldsymbol{\theta}_0)$ are diagonal, it is easy to obtain $\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)$. After simplification, we have the following results:

$$
\begin{aligned}
Var(\hat{\mu}_i) &= \frac{\sigma_i^2}{n_i} \cdot (\delta_0 + \frac{\delta_0^2 \cdot \delta_1^2}{\delta_2 - \delta_0 \cdot \delta_1^2}), \\
Var(\hat{\sigma}_i) &= \frac{\sigma_i^4}{n_i} \cdot \frac{1}{\delta_2 - \delta_0 \cdot \delta_1^2}, \\
Cov(\hat{\mu}_i, \hat{\sigma}_i) &= -\frac{\sigma_i^3}{n_i} \cdot \frac{\delta_0 \cdot \delta_1}{\delta_2 - \delta_0 \cdot \delta_1^2}.
\end{aligned}
$$

For simplicity, in these expressions we replace $\delta_r(\mu_i, \sigma_i)$ with δ_r, $r = 0, 1, 2$, and the detailed expressions for $\delta_r(\mu_i, \sigma_i)$ are given in (B.3), (B.4), and (B.5), respectively.

Proof of Theorem 3.3

Let $\hat{\eta}_i = \hat{\mu}_i - \hat{\mu}_0$, $i = 1, \ldots, G$. From the proof of Theorem 3.2, for $i \neq j$, $\hat{\mu}_i$ and $\hat{\mu}_j$ are independent. Thus, we have

$$
\begin{aligned}
& Cov(\hat{\eta}_i, \hat{\eta}_j) \\
= \; & Cov(\hat{\mu}_i - \hat{\mu}_0, \hat{\mu}_j - \hat{\mu}_0) \\
= \; & Cov(\hat{\mu}_i, \hat{\mu}_j) - Cov(\hat{\mu}_i, \hat{\mu}_0) - Cov(\hat{\mu}_0, \hat{\mu}_j) + Var(\hat{\mu}_0) \\
= \; & Var(\hat{\mu}_0) \\
= \; & \frac{1}{n_0} \cdot \delta(0, 1).
\end{aligned}
$$

Note that the above equation is true for any $i \neq j$, and the result does not depend on the subscripts i and j. This completes the proof of Theorem 3.3.

Appendix C

The Proof of Lemma 4.1

Note that the two-step procedure proposed in Section 4.3.1 by using of a specific treatment, say treatment s, in Step 1, actually rescales the original (unknown) underlying distributions. More specifically, suppose $X_i \sim F(x; \mu_i, \sigma_i)$, $i = 1, \cdots, G$, are the original latent variables of different treatments, and $(\tau_1, \cdots, \tau_{K-1})$ are the original thresholds. We denote the rescaled underlying variables and parameters by $X_i^{(s)} \sim F(x; \mu_i^{(s)}, \sigma_i^{(s)})$, $i = 1, \cdots, G$, and $(\tau_1^{(s)}, \cdots, \tau_{K-1}^{(s)})$, respectively, when the thresholds are determined by treatment s in step 1. The two-step estimation procedure rescales the underlying distributions by performing the following transformation.

$$X_i^{(s)} = h(X_i) = \frac{X_i - \mu_s}{\sigma_s}, \quad i = 1, \cdots, G,$$

$$\tau_k^{(s)} = h(\tau_k) = \frac{\tau_k - \mu_s}{\sigma_s}, \quad k = 1, \cdots, K-1.$$

This reduce to

$$\mu_i^{(s)} = \frac{\mu_i - \mu_s}{\sigma_s}, \quad \sigma_i^{(s)} = \frac{\sigma_i}{\sigma_s}.$$

Since $h(\cdot)$ is a monotonic increasing function, this transformation does not change the ordering of the efficacies of different treatments.

Note that in the above transformation, the cell probabilities, $\pi_{ik}(\boldsymbol{\theta})$, $i = 1, \cdots, G$, $k = 1, \cdots, K$, always keep unchanged. Thus, the test of the dispersion effect by the LRT (4.23) will not be affected by this transformation.

Now, we consider the test of the location effect. Under this transformation, the test statistics can be written as

$$Z_{ij}^{(s)} = \frac{\hat{\mu}_i^{(s)} - \hat{\mu}_j^{(s)}}{\sqrt{\frac{\hat{\sigma}_i^{2(s)}}{n_i}\hat{\delta}^{(s)}(\hat{\mu}_i^{(s)}, \hat{\sigma}_i^{(s)}) + \frac{\hat{\sigma}_j^{2(s)}}{n_j}\hat{\delta}^{(s)}(\hat{\mu}_j^{(s)}, \hat{\sigma}_j^{(s)})}},$$

where $\hat{\delta}^{(s)}(\hat{\mu}_i^{(s)}, \hat{\sigma}_i^{(s)})$ is the estimate of $\delta^{(s)}(\mu_i^{(s)}, \sigma_i^{(s)})$. The expression of $\delta^{(s)}(\mu_i^{(s)}, \sigma_i^{(s)})$ has the same form as $\delta(\mu_i, \sigma_i)$ with τ_k, μ_i, and σ_i in (4.17) replaced by $\tau_k^{(s)}$, $\mu_i^{(s)}$, and $\sigma_i^{(s)}$, respectively. Under mild regularity conditions, the MLEs, $\hat{\mu}_i^{(s)}$, $\hat{\sigma}_i^{(s)}$, and $\hat{\tau}_k^{(s)}$ are consistent estimates of $\mu_i^{(s)}$, $\sigma_i^{(s)}$, and $\tau_k^{(s)}$, respectively. From the formula of Taylor expansion, $Z_{ij}^{(s)}$ can be written as

$$\begin{aligned} Z_{ij}^{(s)} &= \frac{\mu_i^{(s)} - \mu_j^{(s)}}{\sqrt{\frac{\sigma_i^{2(s)}}{n_i}\delta^{(s)}(\mu_i^{(s)}, \sigma_i^{(s)}) + \frac{\sigma_j^{2(s)}}{n_j}\delta^{(s)}(\mu_j^{(s)}, \sigma_j^{(s)})}} + R_n \\ &= \frac{\mu_i - \mu_j}{\sqrt{\frac{\sigma_i^2}{n_i}\delta(\mu_i, \sigma_i) + \frac{\sigma_j^2}{n_j}\delta(\mu_j, \sigma_j)}} + R_n, \end{aligned}$$

where R_n will converge in probability to zero. In the above second equality, we use the result that $\delta^{(s)}(\mu_i^{(s)}, \sigma_i^{(s)}) = \delta(\mu_i, \sigma_i)$. In fact, closer inspection of the expression of $\delta^{(s)}(\mu_i^{(s)}, \sigma_i^{(s)})$ gives that $\delta_0^{(s)}(\mu_i^{(s)}, \sigma_i^{(s)}) = \delta_0(\mu_i, \sigma_i)$, $\delta_1^{(s)}(\mu_i^{(s)}, \sigma_i^{(s)}) = 1/\sigma_s \cdot \delta_1(\mu_i, \sigma_i)$, and $\delta_2^{(s)}(\mu_i^{(s)}, \sigma_i^{(s)}) = 1/\sigma_s^2 \cdot \delta_2(\mu_i, \sigma_i)$. The parameter σ_s will be canceled eventually in the expression of $\delta^{(s)}(\mu_i^{(s)}, \sigma_i^{(s)})$. This result means that, for large sample size, the selection of different treatment to determine the thresholds in step 1 has no effect on the testing results.

Appendix D

The Score Function and Fisher Information Matrix for the Likelihood Function (4.14) and (4.22)

Appendix D-1

We consider the score function and the Fisher information matrix of the likelihood function given in (4.14). We only outline the main results here. A detailed derivation has been given in Appendix B where the underlying distribution is specified as normal distribution. Based on a similar arguments, the following results can be derived when the underlying distribution belongs to the location-scale family.

The score function of the likelihood function (4.14) is

$$\frac{\partial L_2(\boldsymbol{\theta}_0)}{\partial \mu_i} = \sum_{k=1}^{K} \frac{n_{ik}}{\pi_{ik}} \frac{\partial \pi_{ik}}{\partial \mu_i} = -\frac{1}{\sigma_i} \sum_{k=1}^{K} \frac{n_{ik}}{\pi_{ik}} [f_0(\frac{\tau_k - \mu_i}{\sigma_i}) - f_0(\frac{\tau_{k-1} - \mu_i}{\sigma_i})],$$

$$\frac{\partial L_2(\boldsymbol{\theta}_0)}{\partial \sigma_i} = -\frac{1}{\sigma_i^2} \sum_{k=1}^{K} \frac{n_{ik}}{\pi_{ik}} [(\tau_k - \mu_i) f_0(\frac{\tau_k - \mu_i}{\sigma_i}) - (\tau_{k-1} - \mu_i) f_0(\frac{\tau_{k-1} - \mu_i}{\sigma_i})],$$

where π_{ik} is given by (4.15) and $f_0(x)$ is the probability density function of $F_0(x)$.

Under regularity conditions, the expected (Fisher) information matrix is given

by

$$I(\boldsymbol{\theta}_0) = E_{\boldsymbol{\theta}_0}\{\frac{\partial L_2(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0} \cdot \frac{\partial L_2(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'}\} = \begin{bmatrix} A & B \\ B' & C \end{bmatrix},$$

where $\boldsymbol{\theta}_0 = (\mu_1, \cdots, \mu_G, \sigma_1, \cdots, \sigma_G)'$; A, B, and C are $G \times G$ matrices with expressions given as follows.

The diagonal elements of matrix A is given by

$$E[\frac{\partial L_2(\boldsymbol{\theta}_0)}{\partial \mu_i}]^2 = \frac{n_i}{\sigma_i^2}\{\sum_{k=1}^{K} \frac{1}{\pi_{ik}}[f_0(\frac{\tau_k - \mu_i}{\sigma_i}) - f_0(\frac{\tau_{k-1} - \mu_i}{\sigma_i})]^2\} = \frac{n_i}{\sigma_i^2} \cdot \frac{1}{\delta_0(\mu_i, \sigma_i)}.$$

The diagonal elements of matrix B is given by

$$\begin{aligned} & E[\frac{\partial L_2(\boldsymbol{\theta}_0)}{\partial \mu_i} \cdot \frac{\partial L_2(\boldsymbol{\theta}_0)}{\partial \sigma_i}] \\ = & \frac{n_i}{\sigma_i^3}\{\sum_{k=1}^{K} \frac{1}{\pi_{ik}} \cdot [f_0(\frac{\tau_k - \mu_i}{\sigma_i}) - f_0(\frac{\tau_{k-1} - \mu_i}{\sigma_i})] \\ & \cdot[(\tau_k - \mu_i)f_0(\frac{\tau_k - \mu_i}{\sigma_i}) - (\tau_{k-1} - \mu_i)f_0(\frac{\tau_{k-1} - \mu_i}{\sigma_i})]\} \\ = & \frac{n_i}{\sigma_i^3} \cdot \delta_1(\mu_i, \sigma_i). \end{aligned}$$

The diagonal elements of matrix C is given by

$$\begin{aligned} & E[\frac{\partial L_2(\boldsymbol{\theta}_0)}{\partial \sigma_i}]^2 \\ = & \frac{n_i}{\sigma_i^4} \cdot \{\sum_{k=1}^{K} \frac{1}{\pi_{ik}} \cdot [(\tau_k - \mu_i)f_0(\frac{\tau_k - \mu_i}{\sigma_i}) - (\tau_{k-1} - \mu_i)f_0(\frac{\tau_{k-1} - \mu_i}{\sigma_i})]^2\} \\ = & \frac{n_i}{\sigma_i^4} \cdot \delta_2(\mu_i, \sigma_i). \end{aligned}$$

Further calculation shows that the off-diagonal elements of matrices A, B, and C are all equal to 0. So, it is easy to derive the inverse of the expected information matrix. After simplification, we have the following results:

$$\begin{aligned} Var(\hat{\mu}_i) &= \frac{\sigma_i^2}{n_i} \cdot (\delta_0 + \frac{\delta_0^2 \cdot \delta_1^2}{\delta_2 - \delta_0 \cdot \delta_1^2}), \\ Var(\hat{\sigma}_i) &= \frac{\sigma_i^4}{n_i} \cdot \frac{1}{\delta_2 - \delta_0 \cdot \delta_1^2}, \\ Cov(\hat{\mu}_i, \hat{\sigma}_i) &= -\frac{\sigma_i^3}{n_i} \cdot \frac{\delta_0 \cdot \delta_1}{\delta_2 - \delta_0 \cdot \delta_1^2}. \end{aligned}$$

For simplicity, in these expressions we replace $\delta_r(\mu_i, \sigma_i)$ with δ_r, $r = 0, 1, 2$.

Appendix D-2

The score function of the likelihood function (4.22) with the constraint $\sigma_1 = \cdots = \sigma_G = \sigma$ is

$$\frac{\partial \tilde{L}_2(\boldsymbol{\theta}_1)}{\partial \mu_i} = -\frac{1}{\sigma} \sum_{k=1}^{K} \frac{n_{ik}}{\pi_{ik}} [f_0(\frac{\tau_k - \mu_i}{\sigma}) - f_0(\frac{\tau_{k-1} - \mu_i}{\sigma})],$$

$$\frac{\partial \tilde{L}_2(\boldsymbol{\theta}_1)}{\partial \sigma} = -\frac{1}{\sigma^2} \sum_{i=1}^{G} \sum_{k=1}^{K} \frac{n_{ik}}{\pi_{ik}} [(\tau_k - \mu_i) f_0(\frac{\tau_k - \mu_i}{\sigma}) - (\tau_{k-1} - \mu_i) f_0(\frac{\tau_{k-1} - \mu_i}{\sigma})].$$

Under regularity conditions, the expected (Fisher) information matrix is given by

$$I(\boldsymbol{\theta}_1) = E_{\boldsymbol{\theta}_1} \{ \frac{\partial \tilde{L}_2(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1} \cdot \frac{\partial \tilde{L}_2(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1'} \}.$$

The elements of the matrix $I(\boldsymbol{\theta}_1)$ are given by

$$E[\frac{\partial \tilde{L}_2(\boldsymbol{\theta}_1)}{\partial \mu_i}]^2 = \frac{n_i}{\sigma^2} \{ \sum_{k=1}^{K} \frac{1}{\pi_{ik}} [f_0(\frac{\tau_k - \mu_i}{\sigma}) - f_0(\frac{\tau_{k-1} - \mu_i}{\sigma})]^2 \} = \frac{n_i}{\sigma^2} \cdot \frac{1}{\tilde{\delta}_0(\mu_i, \sigma)},$$

$$\begin{aligned} & E[\frac{\partial \tilde{L}_2(\boldsymbol{\theta}_1)}{\partial \mu_i} \cdot \frac{\partial \tilde{L}_2(\boldsymbol{\theta}_1)}{\partial \sigma}] \\ = \; & \frac{n_i}{\sigma^3} \{ \sum_{k=1}^{K} \frac{1}{\pi_{ik}} \cdot [f_0(\frac{\tau_k - \mu_i}{\sigma}) - f_0(\frac{\tau_{k-1} - \mu_i}{\sigma})] \\ & \cdot [(\tau_k - \mu_i) f_0(\frac{\tau_k - \mu_i}{\sigma}) - (\tau_{k-1} - \mu_i) f_0(\frac{\tau_{k-1} - \mu_i}{\sigma})] \} \\ = \; & \frac{n_i}{\sigma^3} \cdot \tilde{\delta}_1(\mu_i, \sigma), \end{aligned}$$

$$\begin{aligned} & E[\frac{\partial \tilde{L}_2(\boldsymbol{\theta}_1)}{\partial \sigma}]^2 \\ = \; & \sum_{i=1}^{G} \frac{n_i}{\sigma^4} \cdot \{ \sum_{k=1}^{K} \frac{1}{\pi_{ik}} \cdot [(\tau_k - \mu_i) f_0(\frac{\tau_k - \mu_i}{\sigma}) - (\tau_{k-1} - \mu_i) f_0(\frac{\tau_{k-1} - \mu_i}{\sigma})]^2 \} \\ = \; & \sum_{i=1}^{G} \frac{n_i}{\sigma^4} \cdot \tilde{\delta}_2(\mu_i, \sigma), \end{aligned}$$

and

$$E[\frac{\partial \tilde{L}_2(\boldsymbol{\theta}_1)}{\partial \mu_i} \cdot \frac{\partial \tilde{L}_2(\boldsymbol{\theta}_1)}{\partial \mu_j}] = 0, \quad 1 \leq i \neq j \leq G.$$

This is the end of the thesis.