A Bayesian Hierarchical Mixture Approach to Model Timing Data with Application to Writing Assessment

Tingxuan Li
A BAYESIAN HIERARCHICAL MIXTURE APPROACH TO MODEL TIMING DATA
WITH APPLICATION TO WRITING ASSESSMENT

By
TINGXUAN LI

A Thesis submitted to the
Department of Educational Psychology and Learning Systems
in partial fulfillment of the
requirements for the degree of
Master of Science

Degree Awarded:
Fall Semester, 2013
Tingxuan Li defended this thesis on June 26, 2013.
The members of the supervisory committee were:

Russell George Almond
Professor Directing Thesis

Young-Suk Kim
Committee Member

Yanyun Yang
Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with university requirements.
To my parents: Li, Wenke & Ru, Caixia, the most wonderful parents in the world.

“When a father gives to his son, both laugh; when a son gives to his father, both cry.”

William Shakespeare

此论文是为了我的父亲李文科和母亲茹彩霞而写的，因为他们是世界上最优秀的家长。

“当父亲给予儿子的时候，他俩都笑了；当儿子回报父亲时，他俩都哭了”

英国作家 莎士比亚
ACKNOWLEDGEMENT

I would like to give a special thanks to Educational Test Services (ETS) for supplying the data used in this thesis, and to my academic advisor—Dr. Russell Almond, who helped me in every stage of my thesis writing.

In addition, I would like to thank the staff at the Statistical Consulting Center in The Florida State University-- in particular, Mr. Michael Rosenthal, who encouraged and mentored me in R programming. I would also like to say thanks to my friends April and Ashton Scott as well as Danisha Baker and Joy McKenzie who encouraged me through their prayers along the way.
<table>
<thead>
<tr>
<th>TABLE OF CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables ................................................................................................................................. vi</td>
</tr>
<tr>
<td>List of Figures ............................................................................................................................... vii</td>
</tr>
<tr>
<td>Abstract ........................................................................................................................................ viii</td>
</tr>
<tr>
<td>CHAPTER ONE – INTRODUCTION ............................................................................................1</td>
</tr>
<tr>
<td>CHAPTER TWO – DATA ..............................................................................................................7</td>
</tr>
<tr>
<td>CHAPTER THREE – MODEL .......................................................................................................9</td>
</tr>
<tr>
<td>CHAPTER FOUR – RESULTS ....................................................................................................17</td>
</tr>
<tr>
<td>CHAPTER FIVE – CONCLUSION/LIMITATION/FUTURE STUDY DIRECTION ..........36</td>
</tr>
<tr>
<td>APPENDICES ...............................................................................................................................39</td>
</tr>
<tr>
<td>A. R CODE FOR DATA EXTRACTION AND DATA CLEANING ...............................................39</td>
</tr>
<tr>
<td>B. R CODE ....................................................................................................................................42</td>
</tr>
<tr>
<td>C. R CODE WITH JAGS ..............................................................................................................48</td>
</tr>
<tr>
<td>D. JAGS MODEL FILE ................................................................................................................51</td>
</tr>
<tr>
<td>E. R CODE FOR CORRELATION ANALYSIS ..........................................................................52</td>
</tr>
<tr>
<td>F. IRB APPROVAL LETTER .......................................................................................................57</td>
</tr>
<tr>
<td>References ......................................................................................................................................58</td>
</tr>
<tr>
<td>Biographical Sketch .......................................................................................................................61</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 4.1 Descriptive statistics for the distribution of lengths of the essays ....................... 25
Table 4.2 Descriptive statistics for the distribution of lengths of the essays
(after data cleaning). .................................................................................................................................. 26
Table 4.3 Summary of estimation results from mixtools.......................................................... 27
Table 4.4 AIC values comparisons for ten essays ................................................................. 28
Table 4.5 Essay8: parameter estimation results ....................................................................... 29
Table 4.6 The two-component hierarchical mixture model parameters for Essay434 ............ 32
Table 4.7 DIC values from the two-component model and the three-component model ....... 32
Table 4.8 Correlation between the timing features and the two-component parameters......... 33
Table 4.9 Correlation between the human scores and the two-component parameters.......... 34
Table 4.10 Correlation for the strand scores and timing features ............................................. 35
LIST OF FIGURES

Figure 3.1 Two-component mixture model in the context of Deane’s multi-layer cognitive writing model ................................................................. 15
Figure 3.2 The plate notation for the proposed model......................................................... 16
Figure 4.1 The number of events for the *BetweenSentence* linguistic context across essays (after data cleaning) ........................................................................................................ 25
Figure 4.2 The box plots of four essays ................................................................................ 26
Figure 4.3 The empirical density plots of four essays .......................................................... 27
Figure 4.4 Essay9: Fitted density (circles) vs.empirical density (solid line) for the two-component model and the three-component model ........................................................................ 28
Figure 4.5 Essay8: Fitted density (circles) vs.empirical density (solid line) for the two-component model and the three-component model ........................................................................ 29
Figure 4.6 Essay434: Traceplots of estimated means ............................................................ 30
Figure 4.7 Essay434: Kernel density plots for two estimated means ..................................... 31
Figure 4.8 Essay434: Autocorrelations for two estimated means ........................................ 31
Figure 4.9 Correlation plots: the variables used are from Table 4.9 ................................. 33
Figure 4.10 Correlation plots: the variables used are from Table 4.10 ............................. 34
Figure 4.11 Correlation Plots: the Variables used are from Table 4.11 ............................ 35
ABSTRACT

Deane (2011) proposed a multi-layer cognitive model of writing. The 2009 Cognitively Based Assessment of, for, and as learning (CBAL) Writing pilot assessment was designed to support that multi-layer model of writing. One of the features of that assessment was that the keystroke activity of students writing essays were logged to computer files. A preliminary analysis (Almond, et al., 2012) developed an algorithm to classify the pauses in writing based on the linguistic context and suggested the distribution for pause events is a mixture of lognormal distributions. This early research was a promising effort to tie the mixture components to the layers in the multi-layer writing model. However, the preliminary analysis with sample size of 68 needed to be repeated with the large data set. Moreover, the data needed to be hierarchically modeled so that the data can fit mixture components when the number of pause events is small per essay. To address these problems, the first part of this thesis fits the large data set (CBAL Writing pilot 2009) to a mixture of lognormal distributions. Then, a distributional analysis was carried out to evaluate the fit of the model to the data. The result showed that the three-component mixture model fits the data better. The second part of the thesis estimates the mixture parameters by using a hierarchical model suggested in the preliminary analysis (Almond, et al., 2012). The hierarchical model is useful when the number of observed pause events is small per essay. The parameters of the hierarchical mixture model were estimated using a Markov Chain Monte Carlo (MCMC) sampler. The results from diagnostic tools of the MCMC sampler showed the sampler converged for both the two- and the three-component models. DIC values showed that the two-component hierarchical mixture model fits the data better.
CHAPTER ONE
INTRODUCTION

Automated scoring systems offer a substantial reduction of cost and increase in the speed of scoring compared to using human essays raters. Many automated measures of writing quality focus on the writing product. Few of them focus on the writing process. The quantification of writing process could help teachers identify students with weak text production skills. The initial CBAL Writing pilot study (Almond, Deane, Quinlan, Wagner, & Sydorenko, 2012) found that data addressing the distribution of pause timing during writing was a mixture of lognormal distributions. This finding corresponds to the multi-layer cognitive writing model, namely, multiple cognitive processes are operating during writing. Such tentative correspondence can be conceptualized as: the first mixture component corresponding to the inscription (i.e., typing words); the second mixture component corresponding to more complex cognitive process (i.e., choosing words).

1.1 Background

ETS’s Cognitively Based Assessment of, for, and as learning (CBAL) is a model for K-12 assessments that serves as both a summative and formative test (Bennett, 2011). The CBAL Writing assessment intends to assess literacy skills (reading, writing, and thinking) together. It requires students to solve a series of rhetorical, conceptual, and linguistic problems (Deane, Odendahl, Quinlan, Fowles, Welsh, & Bivens-Tatum, 2008). The key idea for the test design is that “writing is a socially driven skill that requires the integration of a wide range of specific capabilities” (Deane, 2011, p.44). It describes how Interpretation (reading), Expression (writing), and Deliberation (thinking) are interconnected with one another (the details were described in Figure 1 of Deane, 2011). This multi-layer writing model can be specified in terms of low-to-
high cognitive processes as following: lexical/orthographic, verbal, textual, conceptual and social. Each process is associated with a list of activities that a writer engages in while writing (Deane, 2013).

Knowledge about how students spread their writing time among these layers could allow instructors to give feedback about the writing process to students. Psycholinguists (e.g., writing experts) found that skilled writers showed different patterns than less-skilled writers in terms of the use of time (Flower, 1981; Matsuhashi, 1987). Fluency-related details (such as significant pauses) during the process of writing can be observed in typing behaviors (Deane, 2011). Computers can collect detailed keystroke logs and capture changes in the document composition and revision process. Since CBAL Writing is a computer-based assessment, the information about a student’s writing process—particularly pauses during writing—can be recorded in a non-intrusive way (Almond, et al., 2012).

Almond et al. (2012) identified significant pauses according to eight linguistic contexts (WithinWord, BackSpace, BetweenWord, BetweenSentence, BetweenParagraph, Multiple Backspace, Single Backspace, and Edit) and captured the timing data associated with them. They found that the distribution of pause events within the most common linguistic contexts (WithinWord, BetweenWord) looked like a mixture of lognormal distributions. Other linguistic contexts (e.g., BetweenSentence, BetweenParagraph) had too few observations to fit the mixture model. The majority of essays had only 1 or 2 paragraphs and less than 25 sentences. However, hierarchical mixture models may overcome this limitation. To validate the hierarchical model suggested by Almond et al. (2012), this thesis applies it to timing data collected from the CBAL Writing pilot 2009. Detailed research questions are described at the end of this chapter.
1.2 Literature Review

1.2.1 Time Pause during Writing

Many researchers have speculated that the timing of pauses during the writing process may convey information regarding writing competency. Matsuhashi (1981) found that the process of writing shows different pause patterns for different discourse purposes (reporting, persuasion, and generalization). Van Hell, Verhoeven, and Van Beijsterveldt (2008) found that pauses in the writing process are different for narrative texts and expository texts in both fourth graders and adults. When a writer pauses during the writing process, he/she may think about spelling (lower level) or think about developing an argument (higher level). Perl (1980) found that less skilled writers tend to attend to mechanical concerns, such as checking spelling or fixing a grammatical error, rather than word choice.

1.2.2 Automated Essay Scoring System

Although the studies reviewed above describe what can be done by measuring the pauses in the writing process, most automated essay scoring systems have worked with the final writing product—the essay itself. In the past few decades, at least four automated text analysis systems have been developed (Project Essay Grade, PEG: Page, 1966; e-rater®: Burstein et al, 1998; Intelligent Essay Assessor, IEA: Landauer, Laham, Rehder, & Schreiner, 1997; and Intellimetric: Elliot, 2003). These systems adopted natural language processing techniques to model human raters, particularly relying on linguistic clues to evaluate an essay. For example, topical-specific vocabulary is used to evaluate writing content. Transitional words or clauses such as “in summary”, “conclude”, and “to summarize” are applied to evaluate analytical thinking represented in an essay.
The cost of human scoring includes staff recruitment and training in addition to the effort of reading the essays. The procedure can take huge amounts of time and effort. It is difficult to accomplish the goal of “reporting timely feedback” in writing assessments that use human scoring (Williamson, et al., 2010). In contrast, automated essay scoring systems offer a substantial reduction of cost and increase in the speed of scoring compared to using human raters (Burstein & Chodorow, 1999; Daigon, 1966). The systems score writing more quickly and with lower per subject costs (although the initial cost may be higher). Moreover, the automated scoring system is more consistent in score production, which provides an advantage for educational researchers in conducting longitudinal studies or score equating/scaling across time (Almond, in press).

1.2.3 The Scoring for CBAL Writing

The CBAL Writing assessment was designed to provide both formative and summative feedback. It consists of a series of multiple-choice and short answer questions followed by an extended (45 minutes) essay question. The scoring for the essay questions was originally based on three strands: Strand I—sentence-level control, Strand II—document-level control, and Strand III—critical thinking (Deane, Quinlan, & Kostin, 2011). The Strand II was dropped in the later work (Deane, Quinlan, & Kostin, 2011). The CBAL Writing development team explored automated scoring systems based on the objective that timely feedback is important for formative assessment purposes (Deane, et al., 2008). They found that Strand I scores could be predicted from existing automated scoring systems (e.g., using features from e-rater). This implied that human raters were only required for scoring Strand III—critical thinking. Although e-rater scores of essays have high correlation with the human ratings, both the e-rater and human scores consider only the final product, and not the process used to produce the essay. Formative
feedback about both the product and the process would be useful for teachers to identify students with weak text production skills.

The initial CBAL Writing pilot study (Almond, et al., 2012) showed that the statistical patterns of pauses during the writing were a promising method. For instance, “longer pauses within words appear to be connected to lower-performing writers, possibly due to inefficiencies in their text production skills, while certain editing behaviors are more characteristic of writers producing more highly valued texts” (Deane, 2011, p.43). In the preliminary analysis, the sample contained 68 essays from the 2007 CBAL Writing pilot. They (Almond, et al., 2012) found that the distribution of pause times on log scale was highly leptokurtic. A mixture of lognormal distributions could produce this high kurtosis. The fit of the multi-component mixture model to the data provided support for Deane’s (2011) multi-layer cognitive writing model.

The initial study (Almond, et al., 2012) had two weaknesses: First, the sample size was small (N=68), and results needed to be replicated and confirmed using data from 2009 CBAL Writing pilot. Second, for some linguistic contexts such as BetweenSentence, the numbers of pause events are rare: typically only 5–15 observations occur in an essay. Hierarchical models can borrow strength across essays in estimating the parameters of the mixture component. They can also help estimate components related to the highest level in Deane’s model, which may only be associated with a few events in an event log. This research was developed to address this research need as follows: (a) to divide the event logs from 2009 CBAL Writing pilot into different pause events according to the linguistic contexts defined by Almond et al. (2012), (b) to transform data onto log scale to reduce the skewness.

The first research task uses the larger data set of the 2009 CBAL Writing pilot to replicate the preliminary result provided by Almond et al. (2012). The specific research question
is whether or not the two-component mixture model is adequate to describe the data. Akaike Information Criterion (AIC) statistics and graphical tools (density plots) are used to evaluate the fit of the two-and three-component mixtures.

The second research task is to follow the suggestion that Almond et al. (2012) provided—to apply hierarchical models to the uncommon linguist contexts (in the thesis, *BetweenSentence*) in the Bayesian framework. The specific research question is whether or not the MCMC sampler will converge to a unimodal posterior distribution.
CHAPTER TWO

DATA

The large data set used in this thesis was gathered by Educational Test Services (ETS), the part of the data analyzed in this data set consists of keystroke log files. This chapter has two sections: the first section describes how the data were collected and the second section describes processing of the data.

2.1 Sampling and Participants

CBAL is a suit of assessments that ETS is developing to support both formative and summative assessments of K-12 students (Fu, Chung, & Wise, 2013). The CBAL Writing assessment design combines innovative, computer-based tasks “targeted at measuring higher-order thinking, as well as more elemental questions directed at lower-level, but still essential, component skills” (Bennett, 2011, p.1). Five CBAL Writing pilots were conducted between 2007 and 2009 with four writing genres—(a) persuasive writing, (b) literary writing, (c) argumentation and summarization writing, and (d) informational writing. “These pilots included data from over 4,800 online administrations carried out with eighth grade students” (Bennet, 2011, p.19).

The data that this thesis uses was collected in 2009 from 2,580 eighth grade students from 21 schools in 12 states (Fu, Chung, & Wise, 2009). According to Almond (personal communication, May 10, 2013), although the participating schools were volunteers, care was taken to produce a sample that was balanced demographically (e.g., high vs. low minority enrollment, high vs. low poverty schools). There were four writing prompts: Service Learning, Invasive Plant Species, Ban Ads, and Mango Street. Each student was assigned two out of the four writing prompts. Two prompts analyzed in this thesis were Mango Street (literary writing) and Ban Ads (persuasive writing).
2.2 Data Processing

The initial data collection and processing conducted by ETS followed the steps described in Almond et al. (2012). The assessment delivery system used keystroke log software to capture information about the timing of data. “The logger worked by comparing the current and revised version of the document every 5 milliseconds. If differences were found, then an event would be logged. The logged information consisted of the time of the event, the position of the change, and the added and removed text” (Almond et al., 2012, p. 9).

Pause events classification worked according to what kind of processing the student was doing at the time. First, Almond et al. (2012) defined eight linguistic contexts: *WithinWord, BetweenWord, BetweenSentence, BetweenParagraph, BackSpace, Single BackSpace, Multiple BackSpace,* and *Edit.* Second, the data processing algorithm assigned a pause type based on one of eight linguistic contexts to each event in the event log. Third, a list of pause times was built for each category. Finally, the data processing algorithm resulted in eight different data vectors according to these linguistic contexts for the collected essays. 1,054 essays were collected in the prompt Mango Street and the prompt Ban Ads. The essay log files were processed by the software described in Almond et al. (2012). This thesis analyzed the *WithinWord* and the *BetweenSentence* data vectors.
CHAPTER THREE
MODEL

This chapter provides details about the models used in this research. The first section, notations and computations, provides an overview of the Bayesian framework with Markov chain Monte Carlo (MCMC) algorithm and briefly describes the Expectation-Maximization (EM) algorithm. The second section, finite mixture models, gives a description of discrete finite mixture models and the preliminary analysis of timing data through mixture modeling. The last section describes the plan of implementation of the hierarchical mixture model with the Bayesian data analysis approach.

3.1 Notations and Computations

Recall that this thesis has two research tasks. The first research task is to replicate the preliminary analysis results by using non-hierarchical mixture models where the EM algorithm is used to estimate the parameters. The second research task is to use hierarchical mixture models to estimate the mixture parameters using MCMC algorithm. This section provides descriptions for these algorithms.

“Statistical inference is concerned with drawing conclusions, from numerical data, about quantities that are not observed” (Gelman, Carlin, Stern, & Rubin, 2004, p. 4). This philosophical approach to Bayesian data analysis unfolds in a series of steps. The first step is to construct an appropriate probability specification of the model. In practice, this often involves setting up large models (e.g., many parameters, high dimensional parameter structure, or complicated probability specifications). The second step involves computing (most often simulating random draws from) the posterior distribution of the parameters of interest through Markov chain Monte Carlo (MCMC) algorithms. The inferences are summarized by random
draws from the posterior distribution of the model parameters. The objective of the MCMC algorithm is to obtain a representative sample of the posterior distribution. Hence, convergence means the MCMC sampler explores a posterior distribution to obtain the sample (Gelman & Hill, 2006).

Let $\theta$ denote the parameters of interest, which are unobservable quantities (vectors), and $y$ denote the observed data. Bayesian statistical inference and conclusions about a parameter $\theta$ are made in terms of conditional probability statements. The unobservable quantities $\theta$ are estimated by using probability statements. In setting up models, the joint posterior distribution of parameter $\theta$ and $y$ is denoted $p(y, \theta)$, which is the product of two densities (shown in Equation 1): prior distribution $p(\theta)$, and sampling distribution (is often called the data likelihood) $p(y|\theta)$.

$$p(y, \theta) = \frac{p(\theta) p(y|\theta)}{c},$$

where $c$ is the normalization constant. $c$ does not have to be known when the MCMC sampler is used because the sampler will iteratively find a value which maximizes $p(y, \theta)$.

Unlike MCMC algorithm, EM algorithm is an iterative method which finds a set of parameter values with high posterior probability. When a flat prior is used, it becomes maximum likelihood estimation. The objective of EM algorithm is to obtain point estimates and associated standard errors to allow for population inference. Hence, convergence means that the estimation process is repeated until the difference in log-likelihoods falls below a convergence threshold (McLachlan, & Peel, 2000). Let the log-likelihood for $\theta$ given $y$ denote as:

$$Log L(\theta) = Log p(y, \theta).$$

Let $\theta^{(0)}$ be the initial value for $\theta$, at the $j$-th step of the EM algorithm. Two steps (E step and M step) are performed. E-Step is to calculate:
\[ p(\theta, \theta^{(j)}) = E_{\theta^{(j)}} \{ \text{LogL}(\theta) \mid y \}, \]  

(3)

M-Step is to choose any value \( \theta^{(j+1)} \) that maximizes \( p(\theta, \theta^{(j)}) \), i.e.,

\[ p(\theta^{(j+1)}, \theta^{(j)}) \geq p(\theta, \theta^{(j)}). \]  

(4)

The E and M steps are alternated until the difference,

\[ L(\theta^{(j+1)}) - L(\theta^{(j)}), \]  

(5)

becomes very small (falls below a convergence threshold).

### 3.2 Finite Mixture Model

Mixture models are useful when the observations are taken from two or more different populations (Gelman, et al., 2004). Each population is referred to as a mixture component in a mixture model. In other words, a random variable is drawn from population which consists of \( K \) components (McLachlan & Peel, 2000). The density function (sampling distribution) of the mixture model is hence

\[ f(\gamma_n \mid \theta_k) = \sum_{k=1}^{K} f(\gamma_n \mid \theta_k) \cdot \pi_k \]  

(6)

\( k \) is the index for the mixture component, which can take value between 1 and \( K \).

\( f(\gamma_n \mid \theta_k) \) indicates that the observation \( y_n \) is distributed according to the \( k \)th mixture component with the component parameter \( \theta_k \). Usually, the mixture components are assumed to follow the same parametric family (e.g., all lognormals, all Possions) with different parameter vectors. \( \pi_k \) is the mixing proportion, which is the proportion of the population from the Mixture Component \( k \).

The constraint on \( \pi_k \) is \( \sum_{k=1}^{K} \pi_k = 1 \).

In practice, it is common for the component membership to be unknown. In this research context, it means that it is unknown from which cognitive process the observed pause event \( y_n \) is drawn. The unobserved variable \( z_n \) is called \textit{latent mixture indicator}. Let the vector \( z = \{ z_1, \ldots, z_n \} \)
be a categorical random variable/vector, taking on the value $1, \ldots, K$, with probabilities $\pi_1, \ldots, \pi_K$. The observed pause events $y_n$ are incomplete data unless the associated component labels $z_n$ are given.

One common approach is to fit the mixture models for small values of $k$ ($k=2, 3, \ldots, k=1$ is not considered because it is not a mixture distributions). Almond et al. (2012) did this with the *WithinWord* and the *BetweenWord* pause data from the 2007 CBAL pilot with sample size of 68 students (essays). They found that the two-component mixture of lognormal distributions was adequate for the majority of the essays evaluated.

In mixture of lognormal distributions, random variables—the pause events on log scale $x_n = \log(y_n)$ are drawn from a mixture of normal distributions with the density function:

$$\sum_{k=1}^{K} f(x_n | \mu_k, \sigma_k) \cdot \pi_k. \quad (7)$$

Let $\pi = (\pi_1, \pi_2, \ldots, \pi_K)$ be the mixing proportion, the proportions of the pauses. $\pi_k$ represents the proportion of pauses coming from the $k$th cognitive process. Associated with each pause event $n$, latent mixture indicator $z_n$ takes on value $1, \ldots, K$, with probabilities $\pi_1, \ldots, \pi_K$. If $z_n = k$, then $x_n$ is distributed according to the $k$th mixture component:

$$x_n | z_n = k \sim N(\mu_k, \sigma_k^2). \quad (8)$$

Figure 3.1 illustrates a mixture model with three density graphs. The lowest graph is the kernel density of data vector for an essay where the pause event $x_n$ (the observed pause event on log scale) is distributed according to the two-component mixture model (Verbal cognitive process with the mixing proportion of 0.29 and Lexical cognitive process with the mixing proportions of 0.71). In other words, $x_n$ is drawn either from the Lexical cognitive process (the top density with $\mu_\text{lexical} = 5, \sigma_\text{lexical}^2 = 1$) or the Verbal cognitive process (the middle density with
\( \mu_{\text{verbal}} = 1, \sigma^2_{\text{verbal}} = 0.8 \). A cognitive process that is higher in the cognitive writing model is expected to generate fewer pause events than a process that is lower in the writing model (R. Almond, personal communication, April 10, 2013). In the illustration, the higher level Verbal component has a smaller mixing proportion parameter than the lower level Lexical component.

### 3.3 Proposed Model

In a hierarchical model, the parameters of individual essays are random variables drawn from the defined population (it is often called higher-level population distribution in the Bayesian framework). Hierarchical models offer two potential benefits over the non-hierarchical alternatives. Because the standard error of the mixture component parameters is based on the number of observations from that mixture in the sample, the non-hierarchical mixture does not produce good estimates when the number of observations assigned to one of the components is small. This may happen for one of two reasons: (1) the total sample size \( N_i \) is small (this was the reason that Almond et al. were unable to fit mixtures models for the BetweenSentence and the BetweenParagraph data), or (2) the value of one of the mixing probabilities is small, so that the number of events associated with that component is small. The latter reason is troubling because some of the higher-level cognitive process in the Deane (2011) writing model may appear only rarely. In both cases, the hierarchical mixture model allows the estimation procedure to borrow strength from other essays in calculating the posterior distribution for the parameter.

Denote each essay \( i, i = 1, 2, 3, \ldots, I \). Hence, the observed variable \( Y_i = \{Y_{i1}, \ldots, Y_{iN_i}\} \) is a random vector with \( N_i \) observed pause events nested in each essay \( i \). Log transformation provides \( X_i = \log(Y_i) \), where \( X_i = \{X_{i1}, \ldots, X_{iN_i}\} \). Therefore, all the mixture model parameters are now also indexed by \( i \): \( \mu_{ik} \) is the mean of the \( k \)th component for essay \( i \). \( \sigma^2_{ik} \) is the variance
of the $k$th component for essay $i$. $\mathbf{\pi}_i = \{\pi_{i1}, ..., \pi_{ik}\}$ is the mixing proportion for essay $i$. The hierarchical model assumes that each of these mixture model parameters comes from a population distribution. Thus,

$$\mathbf{\pi}_i \sim \text{Dirichlet} (\alpha_1, ..., \alpha_k)$$

$$\mu_{ik} \sim \text{N} (\mu_{ok}, \tau_{ok})$$

$$\sigma_{ik}^{-2} \sim \text{Scaled Inverse } \chi^2 (\sigma_{ok}^{-2}, \nu_{ok}).$$

The plate notation (Figure 3.2) provides a graphical representation of the Bayesian hierarchical mixture structure. The parameters of interest for essay $i$ are $\theta_i$. $\theta_i$ includes $\mathbf{\pi}_i, \mu_{ik}, \sigma_{ik}^{-2}$. Since little information is known about $\mu_{ok}, \tau_{ok}, \sigma_{ok}^{-2}, \nu_{ok}, \alpha_k$, diffuse priors are assigned for these parameters.

For the non-hierarchical mixture model estimation, the R package mixtools (Benaglia, Chauveau, Hunter, & Young, 2009) was used to fit both the two- and three-component to the $WithinWord$ linguistic context data, to replicate the results of Almond et al. (2012).

Next, for the uncommon linguistic context-- $BetweenSentence$ data, the parameters of the hierarchical model were estimated using Markov Chain Monte Carlo (MCMC) as implemented in JAGS (Plummer, 2012). JAGS output provided the MCMC convergence results.
Figure 3.1 Two-component mixture model in the context of Deane’s multi-layer cognitive writing model
Figure 3.2 The plate notation for the proposed model.
CHAPTER FOUR

RESULTS

This chapter provides the estimation results from the models described in the previous chapter. It is broken into five sections. The first section describes the data cleaning. The second section describes the initial pause distribution, where the data are described and transformed onto log scale. The third section provides the results of non-hierarchical mixture modeling. The fourth section provides the results of hierarchical mixture modeling with the Bayesian data analysis approach, where the MCMC sampler convergence analysis is provided. The fifth section shows correlation analysis results to see how timing features are correlated with other statistics.

4.1 Data Cleaning

Pause data from two writing prompts (Ban Ads and Mango Street) were classified into two different linguistic contexts (WithinWord and BetweenSentence) by using Almond et al. (2012)’s algorithm. The number of the pause counts varied from essay to essay. Table 4.1 shows the descriptive statistics which indicates that a large number of essays were short (fewer than 415 characters, or 11 sentences). In particular, the number of BetweenSentence pause events was extremely small for mixture modeling. In this research context, the simplest model (the two-component mixture of lognormals) has 5 parameters, the least number of model parameters is 5 (the two-component mixture of lognormals), so, trying to estimate 5 parameters using 15 data points would result in very large standard errors.

Therefore, I cleaned the dataset by removing short essays before I ran the mixture modeling. Essays with fewer than 30 WithinWord pause events and essays with fewer than 5 BetweenSentence pause events were deleted. Table 4.2 shows the corresponding descriptive
statistics after the data cleaning. The skewness of the BetweenSentence linguistic context data is reduced after the data cleaning, although it is still skewed (see Figure 4.1).

In the analysis of Ban Ads prompt, the model estimation used samples with the sample size of 963 for the WithinWord linguistic context and the sample size of 708 for the BetweenSentence. In the analysis of Mango Street prompt, the model estimations used samples with the sample sizes of 981 for the WithinWord linguistic context and the sample size of 613 for the BetweenSentence. Appendix A shows the data cleaning code.

4.2 Initial Pause Distribution

A set of box plots was produced to look at the distribution of four linguistic contexts pause events (WithinWord and BetweenWord for each prompt). For illustration purposes, Figure 4.2 shows plots of the distributional shape of four randomly selected Ban Ads essays. As Figure 4.2 indicates, even after transforming the data onto the log scale, the data still have high kurtosis. This shape is a feature of mixture distributions. Namely, the mixture distributions with different variances have higher kurtosis than a normal distribution. The density plots in Figure 4.3 depict the four essay data vectors used in the previous box plots. The features of mixture distributions (heavy tails or high peak) are particularly apparent for the WithinWord linguistic context data. For the BetweenSentence linguistic context, the number of events is not large enough to show detailed features of the distribution.

4.3 The Mixture Modeling

Almond et al. (2012) found that a two-component mixture of lognormal distributions was adequate to fit the WithinWord linguistic context and the BetweenWord linguistic context data. This section shows the results of the non-hierarchical mixture model parameters estimation and compares the fit of the two-and three-component models using the WithinWord data in both
prompts. The one-component lognormal mixture model was not evaluated because the one-component mixture of lognormal distributions is a lognormal distribution, which indicates the pause events are all drawn from one population, in this case, all drawn from one cognitive process. Moreover, the preliminary analysis (Almond et al., 2012) found that the distributions had higher kurtosis than the one-component lognormal mixture model.

An R package mixtools (Benaglia, Chauveau, Hunter, & Young, 2009) estimated the parameters. Then, the number of essays converged to the two-component model and the number of essays converged to the three-component model were counted, the summary is shown in Table 4.3. Next, AIC statistics were calculated based on the log likelihood values that were provided by mixtools:

\[
\text{AIC} = -2 \times \text{Loglikelihood} + 2 \times p. \tag{8}
\]

In Equation 8, \( p \) is the number of parameters in a model. For the two-component mixture model, the value of \( p \) is 5. For the three-component mixture model, the value of \( p \) is 8.

For illustration purpose, ten essays were randomly selected in Ban Ads prompt that converged for both the two- and the three-component solutions to compare AIC statistics (shown in Table 4.4). AIC statistics indicated that for most essays, the fit was close between two models, although AIC statistics favored the three-component model slightly. For example, for Essay9, the density plots (Figure 4.4) show that the lack of fit was not significant for the two-component model while AIC statistics favored its counterpart (three-component model).

Meanwhile, it is worth noting that for some cases, AIC statistics favored the two-component model (shown the last row of Table 4.3). In other words, Table 4.3 shows that the three-component model is favored in all but 30 cases. For example, Essay8 was one of 30 cases where the two-component model was favored. Figure 4.5 provides the density plots of two
models for Essay8. The three-component density showed the misfit compared with its two-component counterpart. The additional component (shown as the spike in the circles in the lower plot) is noise and indicates overfitting the data. This reinforced the idea provided by AIC. Table 4.5 shows that the corresponding estimated parameters for Essay8. The third component had standard deviation close to 0, which indicates that this estimated component did not describe the data very well.

4.4 Hierarchical Mixture Models with a Bayesian Data Analysis Approach

This section provides MCMC convergence analysis from output of Bayesian MCMC analysis in JAGS program. The output shows the details of the sampler convergence for both the two- and the three-component hierarchical mixture models. This thesis used R2jags package (Su & Yajima, 2013), to analyze JAGS output in the R environment. The WithinWord and the BetweenSentence linguistic contexts data collected from both the Ban Ads and the Mango Street prompts were used. This analysis used the following diagnostic tools: (a) $\hat{R}$ measure, (b) traceplot, (c) kernel density plots, and (d) autocorrelations.

$\hat{R}$ is an important measure for MCMC convergence (Gelman, et al., 2004). To calculate $\hat{R}$, multiple MCMC chains are started, often from different starting points (this thesis used two starting from random points). The ratio of within to between chain variance is denoted $\hat{R}$. If the chains mixed well, that is, if the sampler reached the stationary distributions, then, the $\hat{R}$ value should be close to 1.0. In this thesis, most of $\hat{R}$ values ranged from 1.0 to 1.2. Across both writing prompts, both linguistic contexts, and both models (the two- and three-component models), less than 5% of the parameters had $\hat{R}$ values were larger than 1.2. The fact that most $\hat{R}$ values were around 1 indicated that the MCMC chains reached the stationary distributions.
A traceplot graphs the parameter value at iteration $t$ against the iteration number. If the model has converged, the traceplot will move around the mode of the stationary distribution (the $y$-axis). Moreover, since this thesis used two chains, the two chains should lie on top of each other during the iterations. The typical shape of traceplots for the parameters in this thesis showed that the chains appeared to be around the stationary distribution ($y$-axis) mode with very small fluctuations. This was the evidence that the chains reached the stationary distribution (the posterior distribution). For illustration purposes, Figure 4.6 shows the traceplots of two parameters (the means from the two-component hierarchical mixture model estimation of the *BetweenSentence* data of Essay434 in Ban Ads prompt).

Kernel density plots are the empirical density plots of the posterior distributions of the model parameters. If the MCMC sampler is smooth and well-behaved, the kernel density plots should look smooth and unimodal. The parameters kernel density plots in this thesis were all look smooth and unimodal. For illustration purpose, Figure 4.7 depicts the kernel density plots for the two parameters mentioned above (the means from the two-component hierarchical mixture model in the estimation of the *BetweenSentence* data of Essay434 in Ban Ads prompt).

An autocorrelation plot is also an important diagnostic tool of the MCMC sample convergence. When high autocorrelation occurs, it indicates that the draws from the chain in the model are highly correlated. Then, the MCMC sampler will be slow to explore the entire posterior distribution. Figure 4.8 shows the autocorrelation plots of the parameter mentioned above. It indicates that no high autocorrelation was found.

MCMC algorithm based models often use DIC values to compare model performance. Table 4.7 provides the DIC values for both the two-component model and the three-component models across both linguistic contexts data (*WithinWord* and *BetweenSentence*) and across both
writing prompts (Ban Ads and Mango Street). It shows that in the hierarchical mixture models comparison, the DIC values favor the smaller model (the two-component model) for both linguistic contexts data across both writing prompts.

The MCMC sampler provides the statistical inference for each parameter by using the summarized information (mean, standard deviation, percentiles) from the posterior distribution. As summarized in Table 4.6 along with \( \hat{R} \) measure, the estimated results for Essay434 were randomly for illustration purpose. The point most often used to make statistical inference is the mean of the posterior distribution.

### 4.5 Correlation Analysis

The previous methods provided eight analysis results, namely, eight sets of parameters, which were provided by using two linguistic contexts (\textit{BetweenSentence, WithinWord}) data, two writing prompts (Ban Ads, Mango Street), and two models (the two-component mixture of lognormals, the three-component mixture of lognormals). In this section, the correlation analysis provides statistical evidence about (a) the relationship between the parameters and the quality of the final writing product as measured by human raters, and (b) the relationship between the parameters and the other timing features.

The variables used in the correlation analysis were timing features, human scores, estimated mixture parameters. Timing features refer to the one-component mixture of lognormal distributions with two parameters (they are the observed mean and the observed standard deviation). Human scores refer to Strand \textit{I} and Strand \textit{III} scores, which measure the students’ writing product. Since all eight analysis results (the estimated parameters from these eight analyses) showed the same patterns in the correlation analyses. For illustration purpose, parameters presented in this section were from the two-component mixture modeling the
WithinWord linguistic context data collected from Ban Ads writing prompts. They are five estimated parameters for each essay (the mean and standard deviation of low-cognitive component, the mean and standard deviation from high-cognitive component, and mixing proportion).

First, there is strong correlation (see details in Table 4.8) between the estimated mean and the standard deviation of the low-cognitive component and the mean of the one-component mixture model (the observed mean). Table 4.8 indicates that it is possible that the correlations are based on the fluency effect. Namely, for any text production skills (transcription, word choice, grammar usage, or intention reflection) represented by low-to-high cognitive process, writing fluency is a paramount consideration (Deane, 2011). Figure 4.9 is the corresponding correlation plots.

Second, Table 4.9 indicates that there is a correlation between the proportion of pause events in the low-cognitive component and the two strand scores. This may imply that students spend more time on the lexical cognitive process tend to have lower strand scores. In other words, modeling the WithinWord linguist context data may be a measure of dysfluency with orthography (Almond, et al., 2012). Recall that the preliminary analysis finding suggested that “longer pauses within words appear to be connected to lower-performing writers, possibly due to inefficiencies in their text production skills” (Deane, 2011, p.43). Orthography is the lowest cognitive process in Deane’s writing model, if an eighth grade student has too many pauses at the location of typing a word (WithinWord pause data), it is more likely that he/she has spelling difficulty instead of engaging with a higher level cognitive process (e.g., thinking about the word choice). Figure 4.10 is the corresponding correlation plots. Interestingly, the complex relationship between the mixing proportion parameter and the two scores was found in the plot,
the correlation is stronger at the upper bound of the plot. The low scoring students have a wide range of values for mixing proportion parameter.

Third, Table 4.10 indicates that a negative moderate correlation was found between the one-component mixture parameters (the observed mean and the observed standard deviation) and strand scores. This suggests that the correlation measure was the typing speed and text production fluency (Almond, et al., 2012). A strong correlation was found between two strand scores. This indicates “writing trait scores tend to correlate strongly with one another, reflecting a general tendency for all aspects of writing quality to advance together” (Deane, 2011, p. 42). Figure 4.11 is the corresponding correlation plots.
Table 4.1 Descriptive statistics for the distribution of lengths of the essays

<table>
<thead>
<tr>
<th></th>
<th>Ban Ads</th>
<th>Mango Street</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WithinWord</td>
<td>BetweenSentence</td>
</tr>
<tr>
<td>Min</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>The 1st quartile</td>
<td>156.0</td>
<td>4.0</td>
</tr>
<tr>
<td>Median</td>
<td>339.5</td>
<td>10.0</td>
</tr>
<tr>
<td>Mean</td>
<td>413.8</td>
<td>10.3</td>
</tr>
<tr>
<td>The 3rd quartile</td>
<td>598.0</td>
<td>15.0</td>
</tr>
<tr>
<td>Max</td>
<td>1652.0</td>
<td>45.0</td>
</tr>
</tbody>
</table>

Figure 4.1 The number of events for the *BetweenSentence* linguistic context across essays (after data cleaning)
Table 4.2 Descriptive statistics for the distribution of lengths of the essays (after data cleaning)

<table>
<thead>
<tr>
<th></th>
<th>Ban Ads</th>
<th>Mango Street</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Within Word</td>
<td>Between Sentence</td>
</tr>
<tr>
<td>Min</td>
<td>31.0</td>
<td>6.0</td>
</tr>
<tr>
<td>The 1\textsuperscript{st} quartile</td>
<td>281.0</td>
<td>9.0</td>
</tr>
<tr>
<td>Median</td>
<td>532.0</td>
<td>13.0</td>
</tr>
<tr>
<td>Mean</td>
<td>567.8</td>
<td>14.3</td>
</tr>
<tr>
<td>The 3\textsuperscript{rd} quartile</td>
<td>787.0</td>
<td>18.0</td>
</tr>
<tr>
<td>Max</td>
<td>2351.0</td>
<td>45.0</td>
</tr>
</tbody>
</table>

Figure 4.2 The box plots of four essays
Figure 4.3 The empirical density plots of four essays

Table 4.3 Summary of estimation results from \textit{mixtools}

<table>
<thead>
<tr>
<th></th>
<th>Ban Ads</th>
<th>Mango Street</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Essays Analyzed</td>
<td>963</td>
<td>981</td>
</tr>
<tr>
<td>Number of Essays for which</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two-component Model Converged</td>
<td>961</td>
<td>980</td>
</tr>
<tr>
<td>Number of Essays for which</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Three-component Model Converged</td>
<td>948</td>
<td>967</td>
</tr>
<tr>
<td>Number of Essays for which AIC</td>
<td>30</td>
<td>41</td>
</tr>
<tr>
<td>Favored for the Two-component</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 4.4 Essay9: Fitted density (circles) vs. empirical density (solid line) for the two-component model and the three-component model

Note: This is an essay that AIC favors the three-component model

Table 4.4 AIC values comparisons for ten essays

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Three-component</td>
<td>1093.1</td>
<td>939.3</td>
<td>52.3</td>
<td>758.2</td>
<td>1092.6</td>
</tr>
<tr>
<td>Two-component</td>
<td>1087.5</td>
<td>966.9</td>
<td>49.2</td>
<td>758.4</td>
<td>1100.2</td>
</tr>
<tr>
<td>Essay6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Essay7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Essay8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Essay9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Essay10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Three-component</td>
<td>216.4</td>
<td>759.7</td>
<td>233.9</td>
<td>1305.1</td>
<td>1198.8</td>
</tr>
<tr>
<td>Two-component</td>
<td>218.5</td>
<td>758.3</td>
<td>230.9</td>
<td>1315.5</td>
<td>1221.4</td>
</tr>
</tbody>
</table>
Figure 4.5 Essay8: Fitted density (circles) vs. empirical density (solid line) for the two-component model and the three-component model

Note: This is an essay that AIC favors the two-component model

Table 4.5 Essay8: parameter estimation results

<table>
<thead>
<tr>
<th>Three-component Model</th>
<th>Two-component Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\mu_1 = -1.33)</td>
<td>(\mu_1 = -1.35)</td>
</tr>
<tr>
<td>(\mu_2 = -0.47)</td>
<td>(\mu_2 = -0.49)</td>
</tr>
<tr>
<td>(\mu_3 = -0.04)</td>
<td></td>
</tr>
<tr>
<td>(\sigma_1 = 0.31)</td>
<td>(\sigma_1 = 0.30)</td>
</tr>
<tr>
<td>(\sigma_2 = 0.49)</td>
<td>(\sigma_2 = 0.47)</td>
</tr>
<tr>
<td>(\sigma_3 = 0.02)</td>
<td></td>
</tr>
<tr>
<td>(\pi_1 = 0.48)</td>
<td>(\pi_1 = 0.55)</td>
</tr>
<tr>
<td>(\pi_2 = 0.48)</td>
<td>(\pi_2 = 0.45)</td>
</tr>
<tr>
<td>(\pi_3 = 0.04)</td>
<td></td>
</tr>
</tbody>
</table>
Figure 4.6 Essay434: Traceplots of estimated means

Note: The top one corresponds to $\mu_{434,1}$, the bottom one corresponds to $\mu_{434,2}$
Figure 4.7 Essay434: Kernel density plots for two estimated means

Figure 4.8 Essay434: Autocorrelation for two estimated means
Table 4.6 The two-component hierarchical mixture model parameters for Essay434

<table>
<thead>
<tr>
<th>parameter</th>
<th>mean</th>
<th>sd</th>
<th>2.5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>97.5%</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{434,1}$</td>
<td>-0.891</td>
<td>0.252</td>
<td>-1.588</td>
<td>-0.986</td>
<td>-0.733</td>
<td>-0.506</td>
<td>-0.506</td>
<td>1.001</td>
</tr>
<tr>
<td>$\mu_{434,2}$</td>
<td>-0.782</td>
<td>0.369</td>
<td>-1.556</td>
<td>-0.980</td>
<td>-0.757</td>
<td>-0.568</td>
<td>-0.115</td>
<td>1.003</td>
</tr>
<tr>
<td>$\tau_{434,1}$</td>
<td>5.555</td>
<td>2.966</td>
<td>2.280</td>
<td>3.768</td>
<td>4.837</td>
<td>6.387</td>
<td>13.278</td>
<td>1.001</td>
</tr>
<tr>
<td>$\tau_{434,2}$</td>
<td>2.627</td>
<td>1.950</td>
<td>0.455</td>
<td>1.402</td>
<td>2.225</td>
<td>3.281</td>
<td>7.261</td>
<td>1.013</td>
</tr>
<tr>
<td>$\pi_{434,1}$</td>
<td>0.682</td>
<td>0.263</td>
<td>0.105</td>
<td>0.496</td>
<td>0.747</td>
<td>0.909</td>
<td>0.998</td>
<td>1.005</td>
</tr>
<tr>
<td>$\pi_{434,2}$</td>
<td>0.318</td>
<td>0.263</td>
<td>0.002</td>
<td>0.091</td>
<td>0.253</td>
<td>0.504</td>
<td>0.895</td>
<td>1.015</td>
</tr>
</tbody>
</table>

Table 4.7 DIC values from the two-component model and the three-component model

<table>
<thead>
<tr>
<th></th>
<th>Ban Ads</th>
<th>Mango Street</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BetweenSentence</td>
<td>WithinWord</td>
</tr>
<tr>
<td>Two-component</td>
<td>Three-component</td>
<td>Two-component</td>
</tr>
<tr>
<td>63,349.7</td>
<td>149,514.2</td>
<td>196,381.9</td>
</tr>
</tbody>
</table>
Table 4.8 Correlation between the timing features and the two-component parameters

<table>
<thead>
<tr>
<th></th>
<th>$\theta_1$</th>
<th></th>
<th>$\theta_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_0$</td>
<td>0.87*</td>
<td>$\sigma_1$</td>
<td>0.71*</td>
</tr>
<tr>
<td>$\sigma_0$</td>
<td>0.43*</td>
<td>$\mu_2$</td>
<td>0.30*</td>
</tr>
<tr>
<td></td>
<td>0.02</td>
<td>$\pi_2$</td>
<td>-0.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\sigma_2$</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Note: $\mu_0$, $\sigma_0$ are one-component lognormal mixture model parameters. $\theta_1$ is the parameter of the low cognitive component, $\theta_2$ is the parameter of the high cognitive component. * indicates p value less than 0.05

Figure 4.9 Correlation plots: the variables used are from Table 4.9

Note: mean, sd correspond to $\mu_0, \sigma_0$ in Table 4.9. lowpie, highmean, lowmean, lowsd correspond to the mixing proportion of low cognitive component, mean of high cognitive component, mean of low cognitive component, standard deviation of low cognitive component in Table 4.9.
Table 4.9 Correlation between the human scores and the two-component parameters

<table>
<thead>
<tr>
<th></th>
<th>$\theta_1$</th>
<th></th>
<th>$\theta_2$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\mu_1$</td>
<td>$\pi_1$</td>
<td>$\sigma_1$</td>
<td>$\mu_2$</td>
</tr>
<tr>
<td>Strand Score I</td>
<td>-0.26*</td>
<td>-0.22*</td>
<td>-0.21</td>
<td>-0.18</td>
</tr>
<tr>
<td>Strand Score III</td>
<td>-0.25*</td>
<td>-0.20</td>
<td>-0.20</td>
<td>-0.17</td>
</tr>
</tbody>
</table>

Note: $\theta_1$ is the parameter of the low cognitive component, $\theta_2$ is the parameter of the high cognitive component.
* indicates $p$ value less than 0.05.

Figure 4.10 Correlation plots: the variables used are from Table 4.10
Note: Score1, score3 correspond to Strand score I, Strand score III in Table 4.9. lowmean, lowsd, lowpie correspond to $\theta_1$ in Table 4.10. Highmean, highsd correspond to $\theta_2$ in Table 4.10.
Table 4.10 Correlation for the strand scores and timing features

<table>
<thead>
<tr>
<th></th>
<th>$\mu_o$</th>
<th>$\sigma_o$</th>
<th>Strand Score I</th>
<th>Strand Score III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strand Score I</td>
<td>-0.32*</td>
<td>-0.25</td>
<td>1.00</td>
<td>0.83</td>
</tr>
<tr>
<td>Strand Score III</td>
<td>-0.30*</td>
<td>-0.23</td>
<td>0.83</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note: * indicates p value less than 0.05

Figure 4.11 Correlation plots: the variables used are from Table 4.11

Note: score1, score3 correspond to Strand score I, Strand score III in Table 4.11.
CHAPTER FIVE
CONCLUSION/LIMITATION/FUTURE STUDY DIRECTION

5.1 Conclusion

This thesis was a secondary data analysis with the tasks of validating the preliminary analysis of Almond et al. (2012) and estimating the hierarchical mixture parameters based on the preliminary analysis suggestion. This thesis used the larger data set of the 2009 CBAL Writing pilot study to answer the following research questions: (a) Whether or not the two-component mixture model is adequate to describe the data. (b) whether or not the MCMC sampler will converge to unimodal posterior distributions for the hierarchical mixture models.

The answer to the first research question is that: In the non-hierarchical mixture model estimation, the three-component mixture model fit the data better, although both models (the two-and three-component models) fit nearly well. To reach this conclusion, the fit measure (AIC) and graphical tools were used. In the preliminary analysis, the tentative conclusion was that “it seems as if the two-component solution provides an adequate description of the data” (Almond, et al., 2012, p.26). In this thesis, the finding showed that the three-component model fit better than the two-component model in the majority of essays. The difference between the preliminary analysis conclusion and this research conclusion is possibly due to the fact that the preliminary analysis used R package FlexMix (Leisch, 2004).

FlexMix would not form the additional component (in this case, the third component) if the sample size for that additional component is too small (less than 5% of the data). In contrast, the current thesis used the mixtools package which does not have this modeling constraint.

The answer to the second research question is that the MCMC sampler was converged in all these eight analyses. It provided eight sets of estimated parameters for the hierarchal mixture
models. In particularly, the *BetweenSentence* linguistic context data were modeled by borrowing *strength* across essays in the hierarchical mixture models.

### 5.2 Limitation

In this study, the participants knew that the writing scores from CBAL Writing plot would not affected their grades. Therefore, some of students may not have made serious effort. For both writing prompts, 25% of the essays contained less than 4 sentences and less than 156 words. Data cleaning (essay deleting) may have led to deleting essays that were produced by students who produced few sentences or words due to being low-performance students. The latter is problematic because these are the students who need the most help.

Second, mixture models were not identifiable. Arbitrarily defining the low-cognitive component as the one with the smaller observed mean is not the only way to identify the model. Alternatively, this (the component membership) could be estimated instead of assigning arbitrary values.

Third, the assessment design was a 45-minute writing task. The higher cognitive processes in the Deane’s writing model may not appear at all in that time span. Therefore, when a mixture modeling is applied in this research context, the number of mixture component may be limited by the short task.

### 5.3 Future Study Direction

Future studies can investigate the systematic difference. There are two kinds of differences in the timing logs which might be related to student writing, and hence should be studied further. For example, whether or not a systematic difference exists between people for whom the two or three-component model fits better. The low scoring students have a wide range of values for the mixing parameter, whether or not there is difference between them.
Eventually, this statistical analysis could contribute to the development of automated scoring system that incorporates the both writing product and writing process. Before scores based on timing data are presented to students, validity studies should be carried out. For example, ETS is planning to collect a small sample of eye-tracking speed data (R. Almond, personal communication, June 26, 2013). This might yield information about what students are attending to during both long and short pauses. The information could be useful to teachers in helping students become better writers.
APPENDIX A

R CODE FOR DATA EXTRACTION AND DATA CLEANING

INWORD for BanAds STARTS HERE:
```r
databan=read.csv(file="C:/Users/tl11e/Desktop/tina/final data/ban.csv", header=TRUE)
inword.countban=databan$nInWord
clean.countban=ifelse(inword.countban>30,inword.countban,NA)
clean.countban
inwoclean.countban=na.omit(clean.countban)
inword.timeban=list()
for (i in 1:length(BanAdsLogs)){
inword.timeban[[i]]=BanAdsLogs[[i]]@InWord
}
length(inword.timeban)
new=c(inword.timeban)
length(new)
cleannew=new[-c(4, 13, 15, 16, 651, 653, 654, 655, 658, 659, 668, 694, 705, 728, 732, 822, 834, 841, 846, 857, 877, 886, 941, 963, 964, 967, 1002, 1020, 1027, 1032, 1049)]
length(cleannew)
times.inwordban=unlist(cleannew)
length(times.inwordban)
length(inwoclean.countban)
```

BetweenStence for Bands STARTS HERE:
```r
databan=read.csv(file="c:/users/tl11e/Desktop/tina/final data/ban.csv", header=TRUE)
besan.countban=databan$nBSentence
clean.countban=ifelse(besan.countban>15,besan.countban,NA)
clean.countban
beclean.countban=na.omit(clean.countban)
besan.timeban=list()
for (i in 1:length(BanAdsLogs)){
besan.timeban[[i]]=BanAdsLogs[[i]]@Sentence
}
length(besan.timeban)
new=c(besan.timeban)
cleannew=new[-c(2, 3, 4, 20, 22, 23, 25, 26, 27, 28, 29, 3, 1037, 1038, 1039, 1045, 1046, 1048, 1049, 1050, 1051, 1052, 1053, 1054)]
length(cleannew)
length(beclean.countban)
times.senban=unlist(cleannew)
length(times.senban)
eeel=cumsum(c(beclean.countban))
```
inword for Mangostreet STARTS HERE:
datamango=read.csv(file="C:/users/tl11e/Desktop/tina/final data/mango.csv", header=TRUE)
inword.countmango=datamango$nInWord
clean.countmango=ifelse(inword.countmango>30,inword.countmango,NA)
clean.countmango
inwoclean.countmango=na.omit(clean.countmango)
inword.timemango=list()
for (i in 1:length(MangoLogs)){
  inword.timemango[[i]]=MangoLogs[[i]]@InWord
}
length(inword.timemango)
new=c(inword.timemango)
length(new)
cleannew=new[-c(22, 31, 32, 33, 61, 64, 68, 75, 77, 89, 98, 104, 107, 885, 892, 894, 896, 913, 922, 953, 956, 957, 984, 987, 997, 1000, 1032, 1037)]
length(cleannew)
times.inwordmango=unlist(cleannew)
length(times.inwordmango)
length(inwoclean.countmango)

BetweenStence for MANGOSTREET STARTS HERE:
datamango=read.csv(file="C:/users/tl11e/Desktop/tina/final data/mango.csv", header=TRUE)
sesan.countmango=datamango$nBSentence
clean.countmango=ifelse(besan.countmango>15,besan.countmango,NA)
clean.countmango
beclean.countmango=na.omit(clean.countmango)
besan.timemango=list()
for (i in 1:length(MangoLogs)){
besan.timemango[[i]]=MangoLogs[[i]]@Sentence
}
length(besan.timemango)
new=c(besan.timemango)
cleannew=new[-c(2, 3, 4, 54)]
length(beclean.countmango)
times.senmango=unlist(cleannew)
length(times.senmango)
eee=cumsum(c(beclean.countmango))
Counts.inwo=inWordSample$counts
summary(counts)

counts.bese=sentenceSample$counts
summary(counts.bese)

par(mfrow=c(2,2))
pause1=t(inWordSample$times)
pause.data=pause1[,4:7]
boxplot(pause.data, xlab="Essay Id", ylab="Pause Time ", use.cols=TRUE, style="quantile", main="InWord Pause Category ", cex.lab=1.5, cex.axis=1.5, cex.main=1.5, cex.sub=1.5)
pause.logdata=log(pause.data)
boxplot(pause.logdata, xlab="Essay Id", ylab="Pause Time ", use.cols=TRUE, style="quantile", main="InWord Pause Category on Log Scale ", cex.lab=1.5, cex.axis=1.5, cex.main=1.5, cex.sub=1.5)
pause2=t(sentenceSample$times)
pause.sen=pause2[,5:8]
boxplot(pause.sen, xlab="Essay Id", ylab="Pause Time ", use.cols=TRUE, style="quantile", main="BetweenSentence Pause Category ", cex.lab=1.5, cex.axis=1.5, cex.main=1.5, cex.sub=1.5)
pause.logsen=log(pause.sen)
boxplot(pause.logsen, xlab="Essay Id", ylab="Pause Time ", use.cols=TRUE, style="quantile", main="BetweenSentence Pause Category on Log Scale ", cex.lab=1.5, cex.axis=1.5, cex.main=1.5, cex.sub=1.5)
par(mfrow=c(2,1))

density{inWordSample$times}
density{sentenceSample$times}

lines(density{inWordSample$times}, lty=d, col=colors[d])

lines(density{sentenceSample$times}, lty=d, col=colors[d])

legend("topright", legend=c("essay 1", "essay 2", "essay 3", "essay 4"),
          lty=1:ncol(pause.logdata))

legend("topright", legend=c("essay 1", "essay 2", "essay 3", "essay 4"),
          lty=1:ncol(pause.logdata))

lines(density{sentenceSample$times}, lty=d, col=colors[d])

lines(density{sentenceSample$times}, lty=d, col=colors[d])

legend("topright", legend=c("essay 1", "essay 2", "essay 3", "essay 4"),
          lty=1:ncol(pause.logdata))

legend("topright", legend=c("essay 1", "essay 2", "essay 3", "essay 4"),
          lty=1:ncol(pause.logdata))

***Box plot ends here

***Density plot starts here

***InWord pause category, the same four essay vectors used in previous box plot

***write a function so that the density plot for these four vectors are in the same graph representation, note they all have either heavy tails or high peak

***BetweenSentence pause category

***write same function to plot a set of density for the four vectors
cex.lab=1.5, cex.axis=1.5, cex.main=1.5, cex.sub=1.5)
  for (d in 1:ncol(pause.logsen)){
    lines(density(pause.logsen[,d],na.rm=TRUE), lty=d, col=colors[d])
  }
legend("topright", legend=c("essay 1","essay 2","essay 3","essay 4"), lty=1:ncol(pause.logdata))

result.k2=list()
  for (d in 1:ncol(y)){
    result.k2[[d]]=normalmixEM(na.omit(y[,d]), k=2)
  }
vec.k2loglik=rep(NA, ncol(y))
  for (i in 1:length(result.k2)) {
    vec.k2loglik[i]=result.k2[[i]]$loglik
  }
tina.AIC=function(loglik, number.para)
  {-2*loglik+2*number.para}
vec.k2AIC=rep(NA, length(vec.k2loglik))
  for (j in 1:length(vec.k2loglik)) {
    vec.k2AIC[j]=tina.AIC(vec.k2loglik[j], 5)
  }
vec.k2AIC=round(vec.k2AIC, digits=1)

result.k3=list()
  for (d in 1:ncol(y)){
    result.k3[[d]]=normalmixEM(na.omit(y[,d]), k=3, maxit = 5000)
  }
vec.k3loglik=rep(NA, ncol(y))
  for (i in 1:length(result.k3)) {
    vec.k3loglik[i]=result.k3[[i]]$loglik
  }
tina.AIC=function(loglik, number.para)
  {-2*loglik+2*number.para}
vec.k3AIC=rep(NA, length(vec.k3loglik))
  for (j in 1:length(vec.k3loglik)) {
    vec.k3AIC[j]=tina.AIC(vec.k3loglik[j], 5)
  }
vec.k3AIC=round(vec.k3AIC, digits=1)

***Density plot ends here

***calculate AIC for two component model starts here

***y is the pause time in a matrix format

Calculate AIC for two component model ends here

***calculate AIC for three component model starts here
vec.k3AIC=rep(NA,length(vec.k3loglik))
for (j in 1:length(vec.k3loglik)) {
  vec.k3AIC[j]=tina.AIC(vec.k3loglik[j],8)
}
vec.k3AIC=round(vec.k3AIC,digits=1)

TCDF.mixture=function(data.vector,estimated.result)
{
  lambda=estimated.result$lambda
  k=length(lambda)

  new=ecdf(na.omit(y[,1]))
  
  new(na.omit(y[,1]))

  par(mfrow=c(2,1))
  qqplot(TCDF.mixture(na.omit(y[,1]),result.k2[[1]]),
  new(na.omit(y[,1])),cex.lab=1.2,
  cex.axis=1.2, cex.main=1.2, cex.sub=1.2,
  col="dark red", xlab="TCDF",
  ylab="ECDF",main="TCDF vs.ECDF for Two-
  component mixture")
  abline(0,1)

  plot(sort(na.omit(y[,1])),sort(new(na.omit
  (y[,1]))), cex.lab=1.2, cex.axis=1.2,
  cex.main=1.2,
  cex.sub=1.2,type="l",ylim=c(0,1),xlab="the
  first essay vector",ylab="Cumulative
  Distribution Function",col="dark red")
  par(new=TRUE)

  plot(sort(na.omit(y[,1])),sort(TCDF.mixture
  (na.omit(y[,1]),result.k2[[1]])),type="l"
  , col="black",main="CDF from the Data
  ***calculate AIC for
  three component
  model ends here

  ***calculate theoretical
  cumulative
distribution
  function starts here

  ***use the results
  from mixtools and
  built-in function
  pnorm to write a
  TCDF

  ***calculate TCDF
  ends here.

  ***Calculate a data
  vector ECDF starts
  here
vs.CDF from the Model", cex.lab=1.2, cex.axis=1.2, cex.main=1.2, cex.sub=1.2, xlab="the first essay vector", ylab="Cumulative Distribution Function"
legend("bottomright", legend=c("CDF from the data=red", "CDF from the model=black"))
est.density=function(data.vector, estimated.result)
{
lambda=estimated.result$lambda
k=length(lambda)
dnorm.mixture=function(data.vector, specific.component)
{
lambda[specific.component] *dnorm(data.vector, mean=estimated.result$mu[specific.component], sd=estimated.result$sigma[specific.component])
}
dnorms=sapply(1:k, dnorm.mixture, data.vector=data.vector)
return(rowSums(dnorms))
}
par(mfrow=c(1,2))
plot(na.omit(y[,8]), est.density(na.omit(y[,8]), result.k2[[8]]), cex.lab=1.0, cex.axis=1.0, cex.main=1.0, cex.sub=1.0, ylab="Density", ylim=c(0,1), xlab="Essay8 data vector", col="blue", xlim=c(-3,5))
par(new=TRUE)
plot(density(na.omit(y[,8])), cex.lab=1.0, cex.axis=1.0, cex.main=1.0, cex.sub=1.0, xlim=c(-3,5), xlab="Essay8 data vector", ylab="Density", ylim=c(0,1), main="Two-component Fitted Density vs. Empirical Density")
legend("topright", legend=c("fitted density=blue", "empirical density=black"))

"new" hence becomes a function

***two-component cumulative distribution functions starts here

***the CDF from the data

***the CDF from the mixture model

***two CDFs end here

***Fitted density starts here

***use the previously estimated result to write a fitted dens

***Empirical Density vs. fitted density starts here
plot(na.omit(y[,8]), est.density(na.omit(y[,8]), result.k3[[8]]), cex.lab=1.0, cex.axis=1.0, cex.main=1.0, cex.sub=1.0, ylab="Density", ylim=c(0,1), xlab="Essay8 data vector", col="blue", xlim=c(-3,5))

par(new=TRUE)
plot(density(na.omit(y[,8])), cex.lab=1.0, cex.axis=1.0, cex.main=1.0, cex.sub=1.0, xlim=c(-3,5), xlab="Essay8 data vector", ylab="Density", ylim=c(0,1), main="Three-component Fitted Density vs. Empirical Density")

legend("topright", legend=c("fitted density=blue","empirical density=black"))
APPENDIX C
R CODE WITH JAGS

pause=t(sentenceSample$times)
vecpause=c(pause)
newpause=na.omit(vecpause)
newpause=c(newpause)

y1=log(newpause)
N=length(newpause)
compon=rep(NA,N)
I=length(sentenceSample$counts)
offset=cumsum(c(1,sentenceSample$counts))
ind=rep(NA,N)
for (n in 1:N){ind[n]<-sum((offset<=n))}
compon=rep(NA,length(ind))
compon[which.min(newpause)] <- 1
compon[which.max(newpause)] <- 2
set.seed(1234)
final.data=list(y1=y1,ind=ind,
compon=compon, I=I, N=N)

pause.params= c("alpha", "mu.high",
"tau.high","a.high","b.high", "mu.low",
"tau.low", "p")

pause.inits=function(){
  list(alpha=c(0.7,0.3), mu.high=c(6,8),
  tau.high=1/c(12,14)^2,
  a.high=c(1,1.1),b.high=c(1,1.1))
}

***put data on the log scale

***Label smaller value as a draw from the component 1, label bigger value as a draw from the component 2

***monitor the parameters
sen.monitor=jags(data=final.data,
inits=pause.inits, pause.params, n.chains=2,  
n.iter=9000, n.burnin=1000, n.thin=20,  
model.file="C:/Users/tl11e/Desktop/try.bug")

mcout=as.mcmc(sen2log)
mcout=as.matrix(mcout)

par(mfrow=c(3,2))  
plot(density(mcout[,1]),main="Posterior  
Distribution for a.high[1]",cex.lab=1.2,  
cex.axis=1.2, cex.main=1.2, cex.sub=1.2)  
plot(density(mcout[,2]),main="Posterior  
Distribution for a.high[2]",cex.lab=1.2,  
cex.axis=1.2, cex.main=1.2, cex.sub=1.2)  
plot(density(mcout[,3]),main="Posterior  
Distribution for alpha[1]",cex.lab=1.2,  
cex.axis=1.2, cex.main=1.2, cex.sub=1.2)  
plot(density(mcout[,4]),main="Posterior  
Distribution for alpha[2]",cex.lab=1.2,  
cex.axis=1.2, cex.main=1.2, cex.sub=1.2)  
plot(density(mcout[,5]),main="Posterior  
Distribution for b.high[1]",cex.lab=1.2,  
cex.axis=1.2, cex.main=1.2, cex.sub=1.2)  
plot(density(mcout[,6]),main="Posterior  
Distribution for b.high[2]",cex.lab=1.2,  
cex.axis=1.2, cex.main=1.2, cex.sub=1.2)

par(mfrow=c(3,2))  
plot(density(mcout[,7]),main="Posterior  
Distribution for Deviance",cex.lab=1.2,  
cex.axis=1.2, cex.main=1.2, cex.sub=1.2)  
plot(density(mcout[,8]),main="Posterior  
Distribution for mu.high[1]",cex.lab=1.2,  
cex.axis=1.2, cex.main=1.2, cex.sub=1.2)  
plot(density(mcout[,9]),main="Posterior  
Distribution for mu.high[2]",cex.lab=1.2,  
cex.axis=1.2, cex.main=1.2, cex.sub=1.2)  
plot(density(mcout[,10]),main="Posterior

***in order to perform the convergence analysis, MCMC output is needed, and here it is stored in a matrix format.
Distribution for tau.high[1], cex.lab=1.2, cex.axis=1.2, cex.main=1.2, cex.sub=1.2
plot(density(mc$out[,11]), main="Posterior Distribution for tau.high[2]", cex.lab=1.2, cex.axis=1.2, cex.main=1.2, cex.sub=1.2)
APPENDIX D

JAGS MODEL FILE

model{
  for (n in 1:N){
    compon[n]~dcat(p[ind[n],1:2])
    y1[n]~dnorm(mu.low[compon[n],ind[n]],
                tau.low[compon[n],ind[n]])
  }
  for(i in 1:I){
    p[i,1:2] ~ ddirch(alpha[]+.01)
    mu.low[1,i]~dnorm(mu.high[1],tau.high[1])
    mu.low[2,i]~dnorm(mu.high[2],tau.high[2])
    tau.low[1,i]~dgamma(a.high[1]+.01,b.high[1]+.01)
  }
  mu.high[1]~dnorm(0,0.00001)
  mu.high[2]~dnorm(0,0.00001)
  tau.high[1]~dgamma(0.1,0.1)
  tau.high[2]~dgamma(0.1,0.1)
  a.high[1]~dgamma(1.1,1.1)
  a.high[2]~dgamma(1.1,1.1)
  b.high[1]~dgamma(1.1,1.1)
  b.high[2]~dgamma(1.1,1.1)
  alpha[1]~dgamma(1.1,1.1)
  alpha[2]~dgamma(1.1,1.1)
}
APPENDIX E

R CODE FOR CORRELATION ANALYSIS

datamango=read.csv(file="C:/Users/Administrator/Desktop/mango.csv", header=TRUE)
besan.countman=datamango$nInWord
clean.countman=ifelse(besan.countman>30,besan.countman,NA)
clean.countman
beclean.countman=na.omit(clean.countman)
count=c(beclean.countman)
length(count)
mangotime=list()
  for (i in 1:length(MangoLogs)){
    mangotime[[i]]=MangoLogs[[i]]@InWord
  }
new=c(mangotime)
clean.mango=new[-c(22, 31, 32
  699, 700, 701, 710, 780, 812, 857,
  987, 997, 1000, 1032, 1037)]
length(clean.mango)

timedata=c(clean.mango)
y1=unlist(timedata)
y1=log(y1)
N=length(y1)
I=length(count)
offset = cumsum(c(1, count))
ind = rep(NA, N)
for (n in 1:N){ind[n]<-sum((offset<=n))}
compon = rep(NA, length(ind))
compon[which.min(y1)] <- 1
compon[which.max(y1)] <- 2
set.seed(1234)
final.data = list(y1=y1, ind=ind, compon=compon, I=I, N=N)
pause.params = c("p", "mu.low", "tau.low")
pause.inits = function(){
  list(alpha=c(0.5, 0.3), mu.high=c(6, 7), tau.high=1/c(12, 14)^2,
       a.high=c(1, 1.5), b.high=c(2, 1.5))
}
bann = jags(data=final.data, inits=pause.inits, pause.params,
            n.chains=1, n.iter=1000, n.burnin=100,
            n.thin=1, model.file="C:/Users/Administrator/Desktop/try.bug")
mcout = as.mcmc(bann)
mulow = seq(2, by=2, length=981)
mumedi = seq(3, by=2, length=981)

meancom1 = rep(NA, 981)
meancom2 = rep(NA, 981)
for (i in 1:length(mulow))
{
  meancom1[i]=mean(mcout[,mulow[i]])
}
for (i in 1:length(mumedi))
{
  meancom2[i]=mean(mcout[,mumedi[i]])
}

plow=seq(1964,2945,by=1)
plow=seq(2946,3927,by=1)
pcom1=rep(NA,981)
pcom2=rep(NA,981)
for (i in 1:length(plow))
{
pcom1[i]=mean(mcout[,plow[i]])
}

for (i in 1:length(pmedi))
{
pcom2[i]=mean(mcout[,pmedi[i]])
}

taulow=seq(3928,by=2,length=981)

for (i in 1:length(taulow))
{
taucom1[i]=mean(mcout[,taulow[i]])
}

for (i in 1:length(taumedi))
{
{taucom2[i] = mean(mcout[,taumedi[i]])}

mangod = read.csv(file="C:/Users/Administrator/Desktop/mango.csv", header = TRUE)

short.id = mangod$ID

length(short.id)

short.id = short.id[-c(22, 31, 987, 997, 1000, 1032, 1037)]

para = data.frame(row.names=paste(short.id), pcom1=pcom1,
                  pcom2=pcom2, meancom1=meancom1, meancom2=meancom2,
                  taucom1=taucom1, taucom2=taucom2)

strand.score = read.csv(file="C:/Users/Administrator/desktop/scoremango.csv", header=TRUE)

score1 = strand.score$strand1
score3 = strand.score$strand3
long.id = strand.score$ID

strand = data.frame(score1=score1, score3=score3,
                    row.names=paste(long.id))

dim(strand)

common = intersect(names(para), names(strand))
common = intersect(rownames(para), rownames(strand))
length(common)
merge=data.frame(para[common,],strand[common,])

dim(merge)

new=na.omit(merge)

dim(new)

cor(new)
APPENDIX F

IRB APPROVAL LETTER

Office of the Vice President For Research
Human Subjects Committee
Tallahassee, Florida 32306
(850) 644-3672 - FAX (850) 644-4392

APPROVAL MEMORANDUM

Date: 12/17/2012
To: Tingquan Li
Dept.: EDUCATIONAL PSYCHOLOGY AND LEARNING SYSTEMS
From: Thomas L. Jacobson, Chair
Re: Use of Human Subjects in Research

The Parameterization of a Finite Mixture Model for Hierarchical Data Structure in a Writing Task

The application that you submitted to this office in regard to the use of human subjects in the research proposal referenced above has been reviewed by the Human Subjects Committee at its meeting on 12/12/2012.

Your project was approved by the Committee.

The Human Subjects Committee has not evaluated your proposal for scientific merit, except to weigh the risk to the human participants and the aspects of the proposal related to potential risk and benefit. This approval does not replace any departmental or other approvals which may be required.

If you submitted a proposed consent form with your application, the approved stamped consent form is attached to this approval notice. Only the stamped version of the consent form may be used in recruiting research subjects.

If the project has not been completed by 12/11/2013 you must request a renewal of approval for continuation of the project. As a courtesy, a renewal notice will be sent to you prior to your expiration date; however, it is your responsibility as the Principal Investigator to timely request renewal of your approval from the Committee.
REFERENCES


BIOGRAPHICAL SKETCH

Tingxuan Li is an international student in The Florida State University.