CONSTRUCTED RESPONSE DATA ANALYSIS USING STRUCTURAL EQUATION MODELING AND TOPIC MODELING

by

SEOHYUN KIM

(Under the Direction of Zhenqiu Lu and Allan S. Cohen)

ABSTRACT

In education, constructed response (CR) items are increasingly being used in standardized tests. Data obtained from CR items usually include written responses and categorical scores that are used for grading those responses. The primary purpose of this dissertation study is to develop new statistical approaches using structural equation modeling (SEM) and topic modeling for analyzing student written responses to CR items and scores obtained from the responses. SEM is a statistical modeling method used to investigate relationships among observed and latent constructs. Topic modeling is primarily intended to help identify latent structures in textual materials.

This dissertation study consists of three research studies, which are connected by the examination of methodological issues dealing with data from CR items. The first study explores the internal consistency of categorical data having different numbers of response categories. This type of categorical data can often be observed in tests with CR items. This study proposes a SEM approach to nonlinear reliability for tests with items having different numbers of ordered categories. A simulation study evaluating the performance of the proposed approach is presented, and an empirical example is provided to illustrate different reliability coefficients. The second

study introduces a topic model to analyze students' written responses to CR items with associated scores for their responses. The proposed model is designed to detect meaningful homogeneous subgroups with regard to the relationships between topic proportions in examinees' responses and scores. The application of the model is demonstrated through data from student responses to CR items and associated scores on a middle grades test of science inquiry practices. The third study deals with students' written responses collected over multiple time points. This study proposes a new model that combines a topic model and a growth curve model to analyze the texts of answers to CR items collected under a longitudinal study design. The application of the proposed approach is illustrated using real data from middle grades students. A simulation study evaluating the proposed model under several practical testing conditions is also provided.

INDEX WORDS: Constructed response items, Growth curve modeling, Latent Dirichlet allocation, Latent variable modeling, Mixture modeling, Reliability, Structural equation modeling, Topic modeling

CONSTRUCTED RESPONSE DATA ANALYSIS USING STRUCTURAL EQUATION MODELING AND TOPIC MODELING

by

SEOHYUN KIM

B.S., Chonnam National University, Korea, 2008

M.S., Seoul National University, Korea, 2013

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2019

CONSTRUCTED RESPONSE DATA ANALYSIS USING STRUCTURAL EQUATION MODELING AND TOPIC MODELING

by

SEOHYUN KIM

Major Professors:	Zhenqiu Lu
	Allan S. Cohen
Committee:	Nicole Lazar
	Shiyu Wang

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2019

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Constructed Response Items

In educational settings, constructed response (CR) items are widely used in various tests to measure intended purposes (Rodriguez, 2003). CR items ask students to construct their own answers, rather than choosing answers from multiple options. Adding CR items in addition to multiple choice items in an assessment often is done in order to expand the information obtained about students' performance (Pellegrino, Wilson, Koenig, & Beatty, 2014). Large-scale assessments such as PARCC, Smarter balanced assessment, PISA, and NAEP include CR items to measure students' ability (NCES, 2014; OECD, 2014; PARCC, 2016; SBAC, 2016). CR items are also used in many research studies (e.g., Bernholt & Parchmann, 2011; Buxton, Allexsaht-Snider, Aghasaleh, Kayumova, Kim, Choi, & Cohen, 2014; Fulcher, 2012; Johnson, Jenkins, & Jewell, 2005). Data from CR items usually consist of textual responses and categorical scores that were assigned based on the textual responses.

The primary purpose of this dissertation study is to propose new statistical methods for analyzing student written responses to CR items and categorical scores assigned to the responses using structural equation modeling (SEM) and topic modeling.

Reliability for Categorical Data

Reliability for a test indicates the precision of the test and is an essential component of evidence for determining the validity of inferences made from test scores (Kane, 1996). A variety of methods exist for the estimation of reliability of a test with categorical items. In

classical test theory, coefficient alpha (Cronbach, 1951) is commonly used to obtain an internal consistency estimate of reliability. Coefficient alpha is based on the following assumptions: (a) errors in observed item scores are not correlated, and (b) items in a test are essentially τ-equivalent with a single common latent factor, i.e., items in the test have the same true score and the same scale (Novick & Lewis, 1967; Raykov, 1997).

Items scored with ordered categories (e.g., Likert-type items) are common in the social and behavioral sciences (Bollen, 1989; Finney & DiStefano, 2006), particularly in tests with constructed response items. Different approaches to reliability estimation have been proposed for such items. The usual approach is through classical test theory. In this regard, Zumbo, Gadermann and Zeisser (2007) proposed ordinal alpha for ordered categorical variables. Ordinal alpha uses polychoric correlations to measure the relationship between the continuous latent variables underlying the categorical variables and has almost the same assumptions as coefficient alpha: (a) errors in underlying continuous scores are not correlated with each other, and (b) underlying continuous scores are essentially τ-equivalent with a single common latent factor. Lee (2007) also considered the reliability for a test consisting of ordered categorical items. This reliability coefficient focused on observed scores and used a multinomial error model to estimate reliability rather than considering latent variables that underlie the observed categorical scores.

Another approach to estimating reliability is through structural equation modeling (SEM) (Bentler, 2009; Bollen, 1989; Green & Yang, 2009; Miller, 1995; Raykov & Shrout, 2002). The SEM approach introduces a factorial structure of a test into the model (Bollen, 1989; Raykov & Shrout, 2002). It also permits a test to have more than one latent factor making it possible to consider flexible factor structures such as a bi-factor structure. Green and Yang (2009) proposed a nonlinear SEM reliability coefficient for a test with ordinal categorical items within an SEM

framework. For such a test, if the structure of a test is well-specified, the nonlinear SEM approach can provide a more accurate estimate of reliability than a linear SEM approach, which treats the categorical scores as continuous scores (Yang & Green, 2015).

Text Analysis Using Topic Models

An ideal way to conduct research using students' writing is to read each piece individually (Grimmer & Stewart, 2013). However, when there are many documents, this approach is time-consuming and expensive. In computer science, researchers have investigated statistical models to detect latent topics within a large text corpus. There are many variations on topic modeling, but most find the clusters of words that characterize the corpus. The clusters are called topics. Hoffman (1999, 2001) introduced probabilistic latent semantic analysis (pLSA), and Blei, Ng, and Jordan (2003) extended the pLSI model and proposed latent Dirichlet allocation (LDA). The LDA model has been a base model for a variety of topic models.

Topic models have been applied to many fields. Griffiths and Steyvers (2004) used LDA to find scientific topics within the abstracts of papers published in the *Proceedings of the National Academy of Sciences* (PNAS). Erosheva, Fienberg, and Lafferty (2004) used a mixed membership model to find topics within the abstracts and references of papers in PNAS. Paul and Dredze (2011) analyzed health-related Twitter messages using an applied LDA. Lauderdale and Clark (2014) combined LDA and a multidimensional item response theory model to investigate political preferences. In an educational context, Kakkonen, Myller, and Sutinen (2006) used LDA for automatic essay scoring. Chen, Chen, and Xing (2015) analyzed Twitter messages at an academic conference context using LDA to find latent topics. Chen, Yu, Zhang, & Yu (2016) analyzed the journal writings of preservice teachers to investigate latent topics and patterns.

Some topic models have been developed to include additional information. For example, Rosen-Zvi, Griffiths, Steyvers, and Smyth (2004) developed an author-topic model to include the author information of documents. Roberts, Stewart, & Airoldi (2016) introduced a topic model that incorporates covariates to investigate the relationship between the covariates and topics. Blei and McAuliffe (2008) introduced the supervised latent Dirichlet allocation model to jointly model documents and associated labels. There also are topic models that incorporate the time at which documents were created or published. Blei and Lafferty (2006), for example, proposed a dynamic topic model that allows topics to evolve over time. Wang, Blei, and Heckerman (2012) introduced a dynamic topic model for continuous time. Gerrish and Blei (2010) extended the dynamic topic model proposed in Blei and Lafferty (2006) and introduced the document influence model, which allows important documents to more greatly influence future topics than the other documents. Glynn, Tokda, Banks, and Howard (2019) extended the dynamic topic model and proposed dynamic linear topic models that incorporate dynamic topic models. We base development of two new models on the use of topic models to include a mixture model with a supervised LDA model and a topic model with a growth curve model.

Overview of Chapters

Chapter 2 deals with reliability for data having unequal numbers of ordered categories. In order to estimate the reliability of this type of data, we developed an SEM approach to calculate reliability for categorical data where there is an uneven number of categories. Chapter 3 introduces an approach for analyzing student written responses to CR items and their scores on these responses. It incorporates a mixture modeling approach into a supervised topic model to detect latent groups of students that have different relationships between their responses to CR items and the scores of their responses. Chapter 4 introduces an approach to analyze CR

responses collected over multiple time points. We developed a Bayesian hierarchical model that combines a topic model and a latent growth curve model to describe changes in students' use of topics over time. Chapter 5 consists of concluding remarks for the three studies and suggestions for future study.

References

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika, 74*(1), 137-143.

Bernholt, S., & Parchmann, I. (2011). Assessing the complexity of students' knowledge in chemistry. *Chemistry Education Research and Practice, 12*(2), 167-173.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (pp. 113-120). ACM.

Blei, D. M. & McAuliffe, J. D. (2008). Supervised topic models. In *Advances in neural information processing systems* (pp. 121-128).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of machine Learning research, 3*, 993-1022.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Buxton, C., Allexsaht-Snider, M., Aghasaleh, R., Kayumova, S., Kim, S., Choi, Y., & Cohen, A. (2014). Potential benefits of bilingual constructed responses science assessments for emergent bilingual learners. *Double Helix*, *2*(1), 1-21.

Chen, B., Chen, X., & Xing, W. (2015). Twitter Archeology of learning analytics and knowledge conferences. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 340-349). ACM.

Chen, Y., Yu, B., Zhang, X., & Yu, Y. (2016). Topic modeling for evaluating students' reflective writing: a case study of pre-service teachers' journals. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 1-5). ACM.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Erosheva, E., Fienberg, S., & Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences, 101*(suppl 1), 5220-5227.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. *Structural equation modeling: A second course*, 269-314.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly*, *9*(2), 113-132.

Gerrish, S., & Blei, D. M. (2010). A language-based approach to measuring scholarly impact. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (pp. 375-382).

Glynn, C., Tokdar, S. T., Howard, B., & Banks, D. L. (2019). Bayesian Analysis of Dynamic Linear Topic Models. *Bayesian Analysis*, *14*(1), 53-80.

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, *74*(1), 155-167.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228-5235.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 267-297.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 50-57). ACM.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, *42*(1), 177-196.

Johnson, E. S., Jenkins, J. R., & Jewell, M. (2005). Analyzing components of reading on performance assessments: An expanded simple view. *Reading Psychology*, *26*(3), 267-283.

Kakkonen, T., Myller, N., & Sutinen, E. (2006). Applying latent Dirichlet allocation to automatic essay grading. In *Advances in Natural Language Processing* (pp. 110-120). Springer Berlin Heidelberg.

Kane, M. (1996). The precision of measurements. *Applied Measurement in Education*, *9*(4), 355-379.

Lauderdale, B. E., & Clark, T. S. (2014). Scaling politically meaningful dimensions using texts and votes. *American Journal of Political Science*, *58*(3), 754-771.

Lee, W. C. (2007). Multinomial and compound multinomial error models for tests with complex item scoring. *Applied Psychological Measurement*, *31*(4), 255-274.

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling*, 2, 255-273.

NCES (2014). *NAEP technical documentation.* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics (NCES). Retrieved from https://nces.ed.gov/nationsreportcard/tdw/instruments/cog.aspx

Novick, M., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, *32*, 1–13.

OECD. (2014). *PISA 2012 technical report*. Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Partnership for Assessment of Readiness for College and Careers (PARCC) (2016). *Final Technical Report for 2015 Administration*. Retrieved from https://www.state.nj.us/education/assessment/district/PARCCTechReport15.pdf

Paul, M. J., & Dredze, M. (2011). You are what you Tweet: Analyzing Twitter for public health. *ICWSM*, *20*, 265-272.

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (2014). *Developing Assessments for the Next Generation Science Standards*. National Academies Press.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*(2), 173-184.

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural equation modeling*, *9*(2), 195-212.

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, *111*(515), 988-1003.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, *40*(2), 163-184.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence* (pp. 487-494). AUAI Press.

Smarter Balanced Assessment Consortium. (2016) 2013-14 *Technical Report.* Retrieved from https://portal.smarterbalanced.org/library/en/2013-14-technicalreport.pdf

Wang, C., Blei, D., & Heckerman, D. (2012). Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298*.

Yang, Y., & Green, S. B. (2015). Evaluation of structural equation modeling estimates of reliability for scales with ordered categorical items. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *11*(1), 23-34.

Zumbo, B., Gadermann, A., & Zeisser, C. (2007). Ordinal versions of coefficient alpha and theta for Likert scales. *Journal of Modern Applied Statistical Methods*, 6(1), 21-29.

CHAPTER 2

A NONLINEAR APPROACH TO RELIABILITY FOR TESTS WITH ITEMS HAVING DIFFERENT NUMBERS OF ORDERED CATEGORIES[1]

[1] Kim, S., Lu, Z., and Cohen, A.S. Accepted by *Applied Psychological Measurement.* Reprinted here with permission of publisher.

Abstract

This study describes a structural equation modeling (SEM) approach to nonlinear reliability for tests with items having different numbers of ordered categories. A simulation study is provided to investigate the performance of this reliability coefficient, coefficient alpha and population reliability for tests having items with different numbers of ordered categories, a one-factor and a bi-factor structure, and different skewness distributions of test scores. Results indicated the proposed reliability coefficient was close to the population reliability in most conditions. An empirical example was used to illustrate the performance of the different coefficients for a test of items with 2 or 3 ordered categories.

Introduction

Items scored with ordered categories are common in social and behavioral sciences (Bollen, 1989; Finney & DiStefano, 2006). Different approaches to estimation of reliability have been proposed for tests with such items. Coefficient alpha (Cronbach, 1951) is probably the most commonly used to obtain an estimate of test reliability and is taken as a lower bound on the internal consistency of the test (Green, Lissitz, & Mulaik, 1977; Novick & Lewis, 1967; Sijtsma, 2009). It is based on the following assumptions: (a) errors in observed item scores are not correlated, and (b) items in a test are essentially τ-equivalent with a single common latent factor (Novick & Lewis, 1967; Raykov, 1997). Green and Yang (2009) have shown that the value of coefficient alpha for ordered categorical data was the same as the values of model based reliabilities, when items were essentially τ-equivalent, and items had the same number of categories. There are also a number of studies that have investigated the effects of the number of response categories on coefficient alpha (e.g., Bandalos & Enders, 1996; Komorita & Graham, 1965; Lissitz & Green, 1975; Lozano, García-Cueto, & Muñiz, 2008; Weng, 2004). These

studies showed that coefficient alpha generally increases as the number of response categories increase. It is not clear, however, what happens when the number of ordered categories on a test becomes more widely separated, e.g., two categories for some items, four or five categories for other items,

Another approach for estimating reliability is through structural equation modeling (SEM) (Bentler, 2009; Bollen, 1989; Green & Yang, 2009; Miller, 1995; Raykov, 1997; Raykov & Shrout, 2002). The SEM approach introduces a factorial structure of the test into the estimation of reliability. It has been shown to be especially useful for estimating reliability for tests having clusters that group items as in testlets (Cho & Kim, 2015; Green & Yang, 2015; Raykov & Shrout, 2002; Yang & Green, 2010). This could be items sharing the same passage in a reading test (e.g., DeMars, 2006; Li, Bolt, & Fu, 2006; Rijmen, 2010) or items having similar reasoning methods as in a mathematics test (e.g., Cronbach & Shavelson, 2004; DeMars, 2006). In such cases, the tests are assumed to have a single general factor (e.g., reading ability or math ability) and group factors that reflect the clusters. Green and Yang (2009) proposed a nonlinear SEM reliability coefficient for a test with ordinal categorical items within an SEM framework. For such a test, if the structure of the test is well-specified, the nonlinear SEM approach has been found to more accurately estimate reliability than the linear SEM approach, as the linear approach treats categorical scores as continuous (Green & Yang, 2009).

Green and Yang (2009) describe a test consisting of items with the same number of categories. However, tests also often consist of items with different numbers of categories. The PARCC assessment for mathematics (2016) and the Smarter Balanced assessments (2016), for example, both include items that are scored with rubrics that have from two to seven points and two to four points, respectively. Likewise, a test such as the Early Development Instrument

(Janus & Offord, 2007) for measuring children's school readiness also has items scored with four to eight points.

The present study was designed to examine the effects on reliability of summed categorical scored tests when a test has uneven numbers of categories. Below, the nonlinear SEM reliability by Green and Yang (2009) is extended to a model of reliability making it appropriate for tests with items having either the same or different numbers of ordered response categories. We provide a numerical calculation to derive the extended nonlinear SEM reliability and conduct a simulation study to evaluate the performance of this nonlinear SEM reliability. The performance of the resulting nonlinear reliability coefficient is compared to coefficient alpha and population reliability for tests with different numbers of categories compared to tests for which the items had the same number of categories. In addition, an example using real data with different numbers of ordered categories is provided to illustrate the application of the different reliabilities.

Reliability in Classical Test Theory

Suppose there are J items in a test. In classical test theory (CTT), an observed score X_j on item j ($j = 1, \dots, J$) is composed of two components, a true score, T_j, and an error score, ϵ_j:

$$X_j = T_j + \epsilon_j \qquad (2.1)$$

Let X, T and ϵ be the sum of observed scores, of true scores, and of error scores, respectively, across J items. Then $X = \sum_{j=1}^{J} X_j$, $T = \sum_{j=1}^{J} T_j$, $\epsilon = \sum_{j=1}^{J} \epsilon_j$, and $X = T + \epsilon$. The reliability coefficient of a test is defined as

$$\rho = \frac{\sigma_T^2}{\sigma_X^2}, \qquad (2.2)$$

where σ_T^2 is the variance of T, and σ_X^2 is the variance of X (Lord & Novick, 1968). Coefficient

alpha (Cronbach 1951) is used to estimate reliability

$$\alpha = \frac{J}{J-1}\left(\frac{\Sigma_{j\neq k} Cov(X_j, X_k)}{\hat{\sigma}_X{}^2}\right), \qquad (2.3)$$

where $Cov(X_j, X_k)$ is the covariance between observed scores for items j and k, and $\hat{\sigma}_X^2$ is the variance of the observed sum scores over all J items.

Reliability in SEM

In the framework of SEM, item scores can be represented as follows (Green & Yang, 2009; Raykov & Shrout, 2002):

$$X_j^* = \lambda_{1j}\eta_1 + \lambda_{2j}\eta_2 + \cdots + \lambda_{Mj}\eta_M + e_j, \qquad (2.4)$$

where X_j^* is a continuous score for item j, M is the number of latent factors, η_m ($1 \leq m \leq M$) are latent factors weighted by corresponding factor loadings λ_{mj}, and $e_j$is a measurement error term. Parameters in Equation (2.4) are typically estimated by maximum likelihood estimation (MLE) under the assumption that observed variables are continuous and normally distributed (Bollen, 1989). Suppose $X^*$and T are the sum of observed scores and of true scores, respectively. Then $X^* = \sum_{j=1}^{J} X_j^*$ and $T = \sum_{j=1}^{J}\sum_{m=1}^{M}\lambda_{mj}\eta_m$. The linear SEM reliability ρ_{lin} is calculated as the ratio of true sum score variance to observed sum score variance as in Equation (2.2):

$$\rho_{lin} = \frac{\sigma_T^2}{\sigma_{X^*}^2} = \frac{Var(\sum_{j=1}^{J}\sum_{m=1}^{M}\lambda_{mj}\eta_m)}{Var(\sum_{j=1}^{J} X_j^*)}. \qquad (2.5)$$

In this case, ρ_{lin} measures the proportion of observed sum score variance that is attributed to the latent factors, $\eta_1, \eta_2, \ldots, \eta_M$ (Bollen, 1989). This linear SEM reliability, ρ_{lin}, is also the same as coefficient omega (McDonald, 1985; McDonald, 1999).

When the observed data are ordinal categorical, fitting linear SEM models such as Equation (2.4) using the MLE method is not desirable as categorical data violate the assumption

of the MLE method, and the MLE provides inflated chi-square estimates and attenuated factor loadings (Bollen, 1989). To address this problem, we consider the observed categorical scores (X_j) as produced from underlying continuous variables (X_j^*), and that a linear SEM model holds for X_j^* as in Equation (2.4) (Bollen, 1989; Finney & DiStefano, 2006). We assume a nonlinear relationship between X_j and X_j^*:

$$X_j = \begin{cases} C_j - 1, & \text{if} \quad X_j^* > v_{C_j-1} \\ \vdots & \vdots \\ 1, & \text{if} \quad v_1 < X_j^* \leq v_2 \\ 0, & \text{if} \quad X_j^* \leq v_1 \end{cases} \qquad (2.6)$$

where C_j is the number of categories for X_j, and the v_i $(i = 1, 2, \dots, C_j - 1)$ are the category thresholds. This allows for different numbers of ordered score categories for each item j. If X_j^* is less than v_1, X_j is equal to 0, for $v_1 < X_j^* \leq v_2$, X_j is equal to 1, and if X_j^* is above v_{C_j-1}, X_j is equal to $C_j - 1$.

In order to estimate the reliability for the nonlinear measurement model, we use the correlation between two parallel tests:

$$\rho_{X\tilde{X}} = \frac{Cov(X, \tilde{X})}{\sqrt{var(X)var(\tilde{X})}}, \qquad (2.7)$$

where X and $\tilde{X}$ are observed sum scores across J items from two parallel tests, which indicate (a) the same latent factors affect the items from the two tests, and (b) the variance of errors and factor loadings are the same for corresponding items for both tests. For categorical items on parallel tests, the thresholds are the same for corresponding items (Bollen, 1989; Green & Yang, 2009). For example, for item j, the underlying continuous scores from two parallel tests using a confirmatory factor analysis model with M latent factors can be expressed as follows:

$$X_j^* = \lambda_{1j}\eta_1 + \lambda_{2j}\eta_2 + \cdots + \lambda_{Mj}\eta_M + e_j$$
$$\tilde{X}_j^* = \lambda_{1j}\eta_1 + \lambda_{2j}\eta_2 + \cdots + \lambda_{Mj}\eta_M + \tilde{e}_j, \qquad (2.8)$$

where λ_{mj} is a factor loading for factor η_m, e_j and $\tilde{e}_j$ are error terms for X_j^* and $\tilde{X}_j^*$, respectively.

The numerator of $\rho_{X\tilde{X}}$ in Equation (2.7) is a covariance between sum scores and can be presented as a function of covariances between two items from the parallel tests:

$$Cov(X, \tilde{X}) = Cov\left(\sum_{j=1}^{J} X_j\,, \sum_{j'=1}^{J} \tilde{X}_{j'}\right) = \sum_{j=1}^{J}\sum_{j'=1}^{J} Cov(X_j, \tilde{X}_{j'}) \qquad (2.9)$$

Suppose item j has C_j categories, item j' has $C_{j'}$ categories, and the vectors of underlying continuous variables $\boldsymbol{X}^* = \left(X_1^*, \ldots, X_J^*\right)$ and $\tilde{\boldsymbol{X}}^* = \left(\tilde{X}_1^*, \ldots, \tilde{X}_J^*\right)$ follow a multivariate normal distribution with variances of 1. Appendix A shows that $Cov\left(X_j, \tilde{X}_{j'}\right)$ can be represented as

$$Cov\left(X_j, \tilde{X}_{j'}\right) = \sum_{k=1}^{C_j-1}\sum_{l=1}^{C_{j'}-1} \Phi_2\left(v_{j_k}, h_{j'_l};\, \rho_{X_j^* X_{j'}^*}\right) - \sum_{k=1}^{C_j-1} \Phi_1\left(v_{j_k}\right) \sum_{l=1}^{C_{j'}-1} \Phi_1\left(h_{j'_l}\right) \qquad (2.10)$$

where $\{v_{j_1}, \ldots, v_{j_{C_j-1}}\}$ and $\{h_{j'_1}, \ldots, h_{j'_{C_{j'}-1}}\}$ are thresholds for items j and j', respectively, $\Phi_1\left(v_{j_k}\right)$ is the cumulative univariate normal distribution function of threshold v_{j_k}, $\Phi_2\left(v_{j_k}, h_{j'_l}; \rho_{X_j^* X_{j'}^*}\right)$ is the cumulative bivariate normal distribution function of thresholds v_{j_k} and $h_{j'_l}$ with correlation $\rho_{X_j^* X_{j'}^*}$, where $\rho_{X_j^* X_{j'}^*}$ is the correlation between two underlying continuous variables. If X_j^* and $X_{j'}^*$ consist of M latent factors as in Equation (2.8), then $\rho_{X_j^* X_{j'}^*}$ can be represented as

$$\rho_{X_j^* X_{j'}^*} = \sum_{m=1}^{M}\sum_{m'=1}^{M} \lambda_{mj}\lambda_{m'j'}\rho_{\eta_m\eta_{m'}}. \qquad (2.11)$$

where $\rho_{\eta_m \eta_{m'}}$ is the correlation between η_m and $\eta_{m'}$. We specifically indicate this correlation derived from a model as $\rho_{M_{jj'}}$ to avoid confusion.

The denominator of $\rho_{X\tilde{X}}$ in Equation (2.7) is

$$\sqrt{Var(X)Var(\tilde{X})} = Var(X) \qquad (2.12)$$

because X_j and $\tilde{X}_j$ are from the parallel tests, and $Var(X)$, the variance of the summed score X, is equal to the sum of covariances between all pairs of items in the test:

$$Var(X) = Var\left(\sum_{j=1}^{J} X_j\right) = \sum_{j=1}^{J}\sum_{j'=1}^{J} Cov(X_j, X_{j'})$$

$$= \sum_{j=1}^{J}\sum_{j'=1}^{J}\left(\sum_{k=1}^{C_j-1}\sum_{l=1}^{C_{j'}-1} \Phi_2\left(\nu_{j_k}, h_{j'_l};\ \rho_{X_j^* X_{j'}^*}\right) - \sum_{k=1}^{C_j-1} \Phi_1(\nu_{j_k}) \sum_{l=1}^{C_{j'}-1} \Phi_1\left(h_{j'_l}\right)\right) \qquad (2.13)$$

where $\rho_{X_j^* X_{j'}^*}$ is the polychoric correlation of scores on the two observed items, j and j'.

The nonlinear reliability coefficient using the SEM framework is expressed as

$$\rho_{non} = \frac{\sum_{j=1}^{J}\sum_{j'=1}^{J}\left[\sum_{k=1}^{C_j-1}\sum_{l=1}^{C_{j'}-1} \Phi_2\left(\nu_{j_k}, h_{j'_l}; \rho_{M_{jj'}}\right) - \sum_{k=1}^{C_j-1} \Phi_1(\nu_{j_k}) \sum_{l=1}^{C_{j'}-1} \Phi_1\left(h_{j'_l}\right)\right]}{\sum_{j=1}^{J}\sum_{j'=1}^{J}\left[\sum_{k=1}^{C_j-1}\sum_{l=1}^{C_{j'}-1} \Phi_2\left(\nu_{j_k}, h_{j'_l};\ \rho_{X_j^* X_{j'}^*}\right) - \sum_{k=1}^{C_j-1} \Phi_1(\nu_{j_k}) \sum_{l=1}^{C_{j'}-1} \Phi_1\left(h_{j'_l}\right)\right]} \quad (2.14).$$

This can be seen as an extension of Equation 21 in Green and Yang (2009). From equation (2.14), we can estimate the internal consistency reliability of a test consisting of items with different numbers of ordered response categories. This reliability coefficient does not require two parallel tests because the parameters in ρ_{non} can be obtained using a single test.

Simulation Study

A simulation study is presented in this section to investigate the performance of ρ_{non}, the proposed nonlinear SEM reliability coefficient using two types of simulated data sets: data

generated from a one-factor model and data generated from a bifactor model. A bifactor model was considered to investigate the performance of $\rho_{non}$for a test having some factors due to clusters of items such as items sharing the same passage. In this study, we evaluated the performance of ρ_{non} in the context of a test with different numbers of response categories and for a test with the same number of response categories. For the purposes of comparison, coefficient alpha (α) and population reliability coefficients for observed sum scores ($\rho_{X\bar{X}}$) were also presented.

Simulation Study Design

Factor structure. Four factor structures were simulated using one-factor or bifactor models. The four structures are summarized in Table 2.1. Models 1 and 2 are based on a one-factor model and Models 3 and 4 are based on a bifactor model with three group factors.

For Models 3 and 4, the correlations between latent factors were all set to 0. For all models, errors were assumed independent. Factor loadings (λ) were all set to 0.7 for Model 1 and Model 2. For Model 3 and Model 4, factor loadings for the general factor were all 0.7, and two sets of factor loadings for the three group factors were considered: all 0.4 and all 0.6.

Types of tests and numbers of ordered categories. Two types of tests were simulated: One type of test consisted of items with the same number of categories; the other type of test consisted of items with uneven numbers of categories. Tests with either two or five ordered categories (labeled as conditions C2 and C5 in the simulation, respectively) were simulated for the first type. For the second type, a combination of two- and five-category items (labeled as condition C25) was generated to compare with reliability estimates from the first type of tests. In the second type of tests, every third item was simulated to have five ordered categories, and the rest were simulated to have two ordered categories.

Distribution of underlying continuous variables. The distribution of the underlying continuous variables ($X_j^* s$) was generated using a multivariate normal distribution with variances of 1. Covariances between variables were determined by the factor models (Model 1 to Model 4). Suppose the vector of underlying continuous variables $\boldsymbol{X}^* = \left(X_1^*, \ldots, X_J^*\right)$ for J items follows a multivariate normal distribution:

$$\boldsymbol{X}^* \sim N(\mathbf{0}, \boldsymbol{\Sigma}), \qquad (2.15)$$

where $\mathbf{0}$ is a $J \times 1$ vector of 0, and $\boldsymbol{\Sigma}$ is the $J \times J$ covariance matrix of $\boldsymbol{X}^*$ with diagonal elements of 1. To be specific, the underlying continuous score on item j for examinee i, x_{ij}^* was generated to have a variance of one as in Equation (2.16) (Bandalos & Enders, 1996; Bernstein & Teng, 1989; Flora & Curran, 2004; Green & Yang, 2009; Yang & Green, 2015)

$$x_{ij}^* = \boldsymbol{\lambda}_j^T \boldsymbol{\eta}_i + \sqrt{\left(1 - \boldsymbol{\lambda}_j^T var(\boldsymbol{\eta}_i)\boldsymbol{\lambda}_j\right)}\epsilon_{ij}, \qquad (2.16)$$

where $\boldsymbol{\lambda}_j$ is a vector of factor loadings for item j, $\boldsymbol{\eta}_i$ is a vector of latent factors, ϵ_{ij} is an error score, and ϵ_{ij} is uncorrelated with $\boldsymbol{\eta}_i$. The error scores were assumed to have a mean of 0 and a variance of 1. Off diagonal elements in the covariance matrix in Equation (2.15) were determined with Equation (2.16) using the parameters of the relevant model.

Distribution of thresholds. Three sets of thresholds were used to generate the categorical data: (a) normal, (b) moderate skew, and (c) mixed skew. To generate the normally distributed data for two- and five-response categories, the sets of thresholds, {0} and {-1.645, -0.643, 0.643, 1.645}, were used, respectively, to transform the underlying continuous variables. Similarly, for the moderate skew condition, {0.7} and {-0.050, 0.772, 1.341, 1.881} were used to generate ordered variables with two- and five-categories, respectively. In the case of the mixed

skew condition, every third item response was generated with the negative moderate skew thresholds and the rest were generated with positive moderate skew thresholds.

For the five-response category tests, the thresholds used to generate the normal and moderate skew distributions were taken from Muthén and Kaplan (1985). The thresholds for the two-response categories were determined to have the same skewness as the corresponding five-response category condition. The skewness values of datasets from the normal and moderate skew conditions were about 0 and 0.7, respectively.

Data Generation

In total there were 54 conditions: 18 conditions with a one factor model (18 = 2 factor structures × 3 distributions of thresholds × 3 sets of item response categories) and 36 conditions with a bifactor model (36 = 2 factor structures × 2 group factor loadings × 3 distributions of thresholds × 3 sets of item response categories). For each of the conditions, we first generated population data, which consisted of 100,000 observations. From the 100,000 observations, we randomly sampled 500 observations without replacement, and replicated this 100 times. As sample size was not a focus of this study, we sampled 500 observations to avoid estimation problems due to small sample size. In addition, in order to calculate the population reliability, $\rho_{X\tilde{X}}$, which requires sum scores from two parallel tests, we generated a second set of population data. The first and second population data share the same underlying continuous true scores as presented in Equation (2.8). The correlation between the observed sum scores from the parallel tests was treated as the population reliability coefficient.

Data Analysis

A one-factor model and a bifactor model were fit to the corresponding generated data. For each condition, ρ_{non} and α were obtained by using the model parameter estimates and the

generated data, respectively. Model parameters were estimated by the method of weighted least squares with mean and variance adjustment (WLSMV; Muthén & Satorra, 1995), as implemented in Mplus 7.4 (Muthén & Muthén, 1998-2015). Polychoric correlations were also estimated using Mplus 7.4. Based on the parameter estimates, ρ_{non} for each replication was calculated using R (R Core Team, 2017). The R code for ρ_{non} is provided in Appendix B. In addition, α for each replication was obtained by using the R package *psych* (Revelle, 2017).

Simulation Study Results

We first examined whether models converged. Some estimation issues occurred for the bifactor models (Model 3 and Model 4, see Table 2.2), especially when the number of response categories was two and data had mixed skewness. The estimation issues included that a residual covariance matrix was not positive definite or residual variances were negative. In order to make meaningful comparison, the datasets that had no estimation issue were used to calculate ρ_{non} and α.

Table 2.3 presents the means of reliability coefficients for the one-factor models (i.e., Model 1 and Model 2). As presented in Table 2.3, the C2 condition had the lowest $\rho_{X\tilde{X}}$ and ρ_{non}, ranging from 0.77 to 0.90. The C5 condition had the highest $\rho_{X\tilde{X}}$ and ρ_{non}, ranging from 0.86 to 0.94. The two reliability coefficients for the mixed number of categories conditions (i.e., condition C25) ranged from 0.80 to 0.92, and were between those for C2 and C5. The values of ρ_{non} were all close to the values of $\rho_{X\tilde{X}}$, and the standard errors of ρ_{non} were all less than or equal to 0.02. For some conditions, ρ_{non} was .01 higher than $\rho_{X\tilde{X}}$. These results are consistent with the results in Yang and Green (2015) with items having the same number of ordered response categories.

Coefficient α in Table 2.3 had the lowest value, when the category condition was either C2 or C25, and had the highest value for the C5 condition. Coefficient α values under the one factor condition were nearly identical with the corresponding $\rho_{X\tilde{X}}$, when items had the same number of response categories (i.e., conditions C2 and C5). In contrast, α was slightly lower than $\rho_{X\tilde{X}}$, when the test included items with uneven numbers of response categories (i.e., condition C25), and the responses had skewed distributions. When the observed scores were generated from the model with nine items with an uneven number of categories (i.e., Model 1 with C25) and the mixed skew condition, the difference between α and the $\rho_{X\tilde{X}}$ was 0.05. The standard errors of α were also less than or equal to 0.02.

Table 2.4 presents the means of reliability coefficients for the bifactor models (Models 3 and 4). The performance of ρ_{non} was similar to that under the one-factor models in that values for ρ_{non} were close to the corresponding values for $\rho_{X\tilde{X}}$. The differences between these two coefficients were generally less than or equal to .01. As was observed for the one-factor model, the C2 condition had the lowest values of $\rho_{X\tilde{X}}$ and ρ_{non}, ranging from 0.82 to 0.96. The C5 condition had the highest values, ranging from 0.90 to 0.98. Coefficient α had the lowest values when the category condition was either C2 or C25. Under the bifactor model, α values were generally lower than $\rho_{X\tilde{X}}$, but similar to $\rho_{X\tilde{X}}$, when the number of items was 18 and the responses were from the bifactor model with the group factor loadings of 0.4. The differences between α and $\rho_{X\tilde{X}}$ values ranged from 0.01 to 0.14. The standard errors of ρ_{non} and α were all less than or equal to 0.02.

The one-factor models and the bifactor models fit the corresponding data well across all the conditions: the CFI were all higher than .98, and the RMSEA were all smaller than .06 with mean values for each condition, ranging from .01 to .02.

Conclusions from Simulation Study

The values of ρ_{non} were very close to the values of $\rho_{X\tilde{X}}$ across all the conditions. The differences between these two coefficients across all conditions were mostly less than 0.01. ρ_{non} and $\rho_{X\tilde{X}}$ both had the highest values under the C5 condition, and had lowest values under the C2 condition. Both reliability coefficients for the C25 condition were placed between the reliability coefficients for C2 and C5. For some conditions, the values of ρ_{non} were slightly higher than the values of $\rho_{X\tilde{X}}$. Coefficient α was close to or the same as the population reliability in conditions with the same number of categories from the one-factor model, but tended to be lower than $\rho_{X\tilde{X}}$ when items had the mixed numbers of response categories or when data were generated from the bifactor model.

Illustration with Empirical Data

In this section, we illustrate the performance of the nonlinear SEM reliability, and compare it with that of coefficient alpha on a set of real data. The data were from an NSF-funded project focusing on teaching science inquiry practices for students in Grades 6 to 10.

Test

The test consisted of 27 constructed response items designed to measure students' use of academic language and understanding of science inquiry practices. Seventeen of the 27 items were scored with a 2-point rubric; the remaining 10 items were scored with a 3-point rubric. Ten items had skewness larger than 0.7, and two items had skewness less than -0.7. Descriptive statistics for each item are provided in Table 2.5.

Sample

The sample consisted of 906 students across grades 6 to 10: 260 (29.05%) students were in the 6th grade, 209 (23.35%) students were in the 7th grade, 222 (24.80%) students were in the

8th grade, 80 (8.94%) students were in the 9th grade, and 124 (13.85%) students were in the 10th grade. There were 11 students for whom grade information was not available.

Results

A one-factor model and a bifactor model were fit to the data. The one-factor solution did not provide good model fit: $\chi^2(324)$ = 4662.22, $\chi^2(324)/(df = 324) = 14.39$, CFI=.93, and RMSEA =.12. For this solution, ρ_{non} was 1.04, which is an unacceptable value for reliability. Also, this one-factor solution could not be used for estimating reliability because the model did not fit the data. The bifactor model with five group factors was next fit to the data. This solution fit the data better: $\chi^2(297) = 841.540$, $\chi^2(297)/(df = 297) = 2.83$, CFI=.991, and RMSEA=.045. The five group factors were consistent with the construction of the test as each factor appeared to represent one of the five science inquiry practices measured by the test. For this bifactor model solution, the $\rho_{non} = .96$ and coefficient alpha was .92.

Discussion

In this study, a generalized nonlinear SEM reliability coefficient was designed to provide internal consistency reliability estimates for tests with items scored with equal or unequal numbers of ordered categories. A simulation study evaluated the performance of this coefficient compared to coefficient alpha and population reliability for observed sum scores. Results indicated that the nonlinear SEM reliability estimates and the population reliability values were close across all the conditions. The nonlinear SEM reliability coefficient and the population reliability coefficient always had the lowest values under the two-response-category condition and had the highest values under the five-response-category condition. The two reliability coefficients for items with mixed numbers of response categories fell in-between. This performance can be expected as having more response categories indicates more information.

Further, items with a mixed number of categories might be assumed to have more information than those having the smallest number of categories among the mixed categories and also to have less information than those having the largest number of categories. The nonlinear SEM reliability successfully captured this tendency by estimating reliability close to the population reliability.

In general, coefficient alpha values were close to the corresponding population reliability coefficient for simulation conditions under the one-factor model and for equal numbers of categories (i.e., conditions C2 and C5). In addition, continuous scores that underlay the observed categorical scores were essentially τ-equivalent under the one-factor model. In generating tests with items with different numbers of categories, different thresholds were applied to the underlying continuous scores of the items to transform the data into categorical scores. Results suggest that this may have resulted in tests which were no longer essentially τ-equivalent. In this case, under the one-factor model and with items having different numbers of categories, coefficient alpha tended to be lower than the population reliability, especially when the scores were skewed. Coefficient alpha was also lower than the corresponding population reliability under the bifactor model conditions (Model 3 and Model 4). These results were expected as the items generated from the bifactor model were no longer essentially τ-equivalent. Cronbach (1951) also noted that when a test has a general factor related to all items with group factors related to part of the items, coefficient alpha is slightly higher than the proportion of variance due to the general factor and lower than the population reliability, which is proportion of variance due to all common factors. In this study, however, when the number of items was 18 and the responses were generated with relatively low group factor loadings (in this simulation

study, 0.4), coefficient alpha, the nonlinear SEM reliability, and the population reliability were similar.

It should be noted that the simulation study was conducted under the conditions of correctly specified factor models and with the assumption of multivariate normality of the latent variables. The empirical data analysis showed the impact of incorrect model specification on the nonlinear SEM reliability estimate. The model fit index showed poor fit for the unidimensional model. Further, the covariance matrix for this model deviated from the sample covariance matrix resulting in reliability estimates larger than 1. For the multivariate normality of latent variables assumption, however, Yang & Green (2015) found that the nonlinear SEM reliability was robust to the modest violation of normality assumption under the condition of the same number of response categories.

References

Bandalos, D. L., & Enders, C. K. (1996). The effects of nonnormality and number of response categories on reliability. *Applied Measurement in Education, 9*(2), 151-160.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika, 74*(1), 137-143.

Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105*(3), 467-477.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods, 18*(2), 207-230.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and psychological measurement, 64*(3), 391-418.

DeMars, C. E. (2006). Application of the Bi-Factor multidimensional item response theory model to Testlet-Based tests. *Journal of educational measurement, 43*(2), 145-168.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. *Structural equation modeling: A second course*, 269-314.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological methods, 9*(4), 466-491.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality1. *Educational and Psychological Measurement, 37*(4), 827-838.

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*(1), 155-167.

Janus, M., & Offord, D. R. (2007). Development and psychometric properties of the Early Development Instrument (EDI): A measure of children's school readiness. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *39*(1), 1-22.

Komorita, S. S., & Graham, W. K. (1965). Number of scale points and the reliability of scales. *Educational and Psychological Measurement*, *25*(4), 987-995.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30*(1), 3-21.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology, 60*(1), 10-13.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *4*(2), 73-79.

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale, NJ: Erlbaum.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling*, 2, 255-273.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*(2), 171-189.

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus user's guide (7th edn)*. Los Angeles, CA: Muthén & Muthén.

Muthén, B. O., & Satorra, A. (1995). Technical aspects of Muthén's LISCOMP approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika*, *60*(4), 489-503.

Novick, M., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32*, 1–13.

Partnership for Assessment of Readiness for College and Careers (PARCC) (2016). *Final Technical Report for 2015 Administration*. Retrieved from https://www.state.nj.us/education/assessment/district/PARCCTechReport15.pdf

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*(2), 173-184.

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural equation modeling*, *9*(2), 195-212.

Revelle, W. R. (2017). psych: Procedures for personality and psychological research. (R Package Version 1.8.4). Retrieved from https://CRAN.R-project.org/package=psych

Rijmen, F. (2010). Formal Relations and an Empirical Comparison among the Bi-Factor, the Testlet, and a Second-Order Multidimensional IRT Model. *Journal of Educational Measurement*, *47*(3), 361-372.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107-120.

Smarter Balanced Assessment Consortium. (2016) 2013-14 *Technical Report.* Retrieved from https://portal.smarterbalanced.org/library/en/2013-14-technicalreport.pdf

Weng, L. J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, *64*(6), 956-972.

Yang, Y., & Green, S. B. (2010). A note on structural equation modeling estimates of reliability. *Structural Equation Modeling*, *17*(1), 66-81.

Yang, Y., & Green, S. B. (2015). Evaluation of structural equation modeling estimates of reliability for scales with ordered categorical items. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *11*(1), 23-34.

CHAPTER 3

EXPLORING STUDENTS' RESPONSES TO CONSTRUCTED RESPONSE ITEMS WITH A SUPERVISED TOPIC MODEL FOR HETEROGENEOUS POPULATIONS[2]

[2] Kim, S., Lu, Z., and Cohen, A.S. Submitted to *Journal of Educational and Behavioral Statistics*, 01/25/19

Abstract

Text data are often paired with numbers, such as scores on essays, on educational tests. Topic modeling is a technique for detecting the latent topics in a collection of documents and has been widely used to analyze texts in a variety of areas. Supervised latent Dirichlet allocation (sLDA) is a topic model that jointly models text data and related labels, such as could occur with students' answers to constructed response (CR) items and their rubric-based scores. In this study, we introduce the finite mixture supervised latent Dirichlet allocation (MixSLDA) model, which extends the sLDA to detect subgroups that are homogeneous in their relationship between topic proportions in student answers and the scores of their answers. This approach is especially useful for situations in which the examinee population is not homogeneous. The application of the model is demonstrated using data from student responses to CR items and associated scores. A simulation study is also presented that evaluates the performance of the MixSLDA under practical testing conditions.

Introduction

Text data are very common in education as many tests or instruments include items that require students' written responses, such as standardized achievement tests (e.g., NAEP, TIMSS) and various research studies (e.g., Andrade, Du, & Mycek, 2010; Furtak & Ruiz-Primo, 2008). These kinds of text data are often paired with numbers, such as essays and essay scores as students' written responses are usually graded for reporting or further analysis. One statistical method that has been actively studied for analyzing text data is topic modeling (Blei, 2012). Topic modeling is a family of statistical models for investigating latent topics from a collection of documents. In this study, we develop a new approach using topic modeling combined with a mixture modeling framework to jointly model text data and associated scores.

In the topic modeling framework, it is assumed that each document in a corpus represents a mixture of topics. One of the simpler and most popular of the topic models is latent Dirichlet allocation (LDA; Blei, Ng, and Jordan, 2003). LDA has been applied in a variety of contexts. For example, Bisgin, Liu, Fang, Xu, and Tong (2011) used LDA to identify topics in drug labels and group drugs using the topics. Guo, Barnes, and Jia (2017) analyzed hotel reviews using LDA to identify key dimensions underlying the reviews. Prier, Smith, Giraud-Carrier, and Hanson (2011) analyzed Twitter messages using LDA to discover health-related trends, especially tobacco use. In education, Chen, Yu, Zhang, and Yu, (2016) analyzed pre-service teachers' journals using LDA to detect the latent thematic structure discussed in students' journals. Kim, Kwak, Cardozo-Gaibisso, Buxton, and Cohen (2017) analyzed middle school students' responses to constructed response items using LDA to detect the latent topic structure in students' answers.

Text data are often paired with additional information, such as essays with essay scores or articles with authors' names. Topic models have been developed to use this additional information to further understand the latent thematic structure in a corpus. For example, Rosen-Zvi, Griffiths, Steyvers, and Smyth (2004) developed an author-topic model by extending LDA to include the author information of documents. Roberts, Stewart, & Airoldi (2016) introduced a topic model that incorporates covariates to investigate the relationship between the covariates and topics. Lacoste-Julien, Sha, and Jordan (2008) introduced the discriminatively trained LDA, which utilizes categorical information associated with each document in modeling topic proportions. Mimno and McCallum (2008) developed the Dirichlet-multinomial regression topic model, which models the parameter of the prior distribution of topic proportions using additional data associated with documents, and Blei and McAuliffe (2008) introduced the supervised LDA

(sLDA) model to jointly model documents and related labels (e.g., scores on essays) to find latent topics that predict the labels of documents.

The focus of this study is on data from constructed response items. Students' responses to constructed response items are usually scored by raters, so the responses are generally associated with scores. Investigating relationships between students' written responses and their scores would be useful for understanding how students obtain scores. In this study, we introduce a new approach that incorporates mixture modeling into sLDA in order to detect latent groups of students that have different relationships between their responses to constructed response items and the scores of their responses. In the proposed model, students' scores are represented by a finite set of regression equations, each of which corresponds to a unique relationship between topic proportions and scores. In education, student populations are often assumed to have subpopulations because of diversity in their demographics or performance or both. In mixture models for example, the presence of score distributions with multiple modes suggests the existence of latent subgroups (McLachlan & Peel, 2000). This kind of multimodal distribution can be observed in distributions of students' essay scores due in part to variability in students' writing. Identifying latent subgroups that share the same relationship between written responses and associated scores can thus add to the information about students' performance on the assessment.

This study begins by introducing LDA, followed by a description of sLDA, and finally the proposed mixture sLDA (MixSLDA). Next, an empirical data analysis is described to motivate the use of the MixSLDA with students' responses to constructed response items and associated scores. A simulation study is then presented to evaluate the performance of the MixSLDA under various conditions.

Models

In this section, we describe latent Dirichlet allocation (LDA), supervised latent Dirichlet allocation (sLDA), and the proposed finite mixture supervised latent Dirichlet allocation (MixSLDA).

Latent Dirichlet Allocation (LDA)

Latent Dirichlet allocation (LDA; Blei, Ng, & Jordan, 2003) is a hierarchical Bayesian model that detects latent topics from a set of documents. Suppose there exists a corpus consisting of D documents, with each document i ($i = 1, \dots, D$) represented as a vector of words, $\boldsymbol{w}_i = \left(w_{i,1}, w_{i,2}, \dots, w_{i,N_i}\right)^T$, where N_i is the number of words in document i. Suppose the corpus has K latent topics, each of which is indexed by k, and there are a total of V unique words in the corpus, with each indexed by v. Here, each topic is presented as a multinomial distribution (γ_k) over V unique words, where γ_k is a $V \times 1$ vector of probabilities. Each document is assumed to be composed of a mixture of topics, and each word ($w_{i,n}$) in a document has a topic indicator, $z_{i,n}$, indicating the topic from which this word came. The topic assignment is determined by a topic proportion vector θ_i that is associated with each document. LDA assumes that each word in a document is randomly drawn from a corresponding topic. In other words, the order of words in the document does not matter. This is referred to as the bag-of-words assumption (Blei, 2012). In addition, topic proportions (θ_i) and word probabilities for each topic (γ_k) are assume to follow a Dirichlet distribution.

Supervised Latent Dirichlet Allocation (sLDA)

LDA is an unsupervised model in that no external information beyond the text is used to guide the analysis. Supervised latent Dirichlet allocation (sLDA; Blei & McAuliffe, 2008), on the other hand, jointly models documents and labels, where labels can include such things as

scores on the constructed response items. This joint model detects latent topics that predict the labels of documents. The sLDA model is constructed by adding an additional step to the LDA model in order to investigate the relationship between topic proportions and a label variable.

Suppose that y_i is a continuous label variable. In sLDA, the relationship between mean topic proportions and the label variable is described using a linear regression as follows:

$$y_i|\boldsymbol{z}_i, \boldsymbol{b}, \sigma^2 \sim N(\boldsymbol{b}^T\bar{\boldsymbol{z}}_i, \sigma^2), \qquad (3.1)$$

where $\boldsymbol{z}_i$ is a vector of topic assignments for document i, which is presented as $\boldsymbol{z}_i = \left(z_{i,1}, z_{i,2}, \dots, z_{i,N_i}\right)^T$, $\boldsymbol{b}$ is a vector of regression coefficients on $\bar{\boldsymbol{z}}_i = \frac{1}{N_i}\sum_{n=1}^{N_i}\tilde{\boldsymbol{z}}_{i,n}$, in which $\tilde{\boldsymbol{z}}_{i,n} = \left(z_{i,n,1}, z_{i,n,2}, \dots z_{i,n,K}\right)^T$ is a $K \times 1$ vector with $z_{i,n,k} = 1$ if $z_{i,n} = k$, and $z_{i,n,k} = 0$, otherwise, and σ^2 is a residual variance of the regression. The vector of regression coefficients $\boldsymbol{b}$ shows how topics are connected with labels. There is no intercept term in the model because the sum of $\bar{\boldsymbol{z}}_i$ components is always 1.

Finite Mixture Supervised Latent Dirichlet Allocation (MixSLDA)

In this study, we limit our interest to continuous types of labels associated with documents (i.e., continuous test scores). The sLDA model is extended here for situations in which the population consists of two or more latent classes, each of which is assumed to have an unique relationship between the topics and the label. In the finite mixture supervised latent Dirichlet allocation (MixSLDA), each document is assumed to belong to one of the latent classes (g). Each latent class has a unique relationship between the topic proportions and the label variable by having distinct regression coefficients and a residual variance. In MixSLDA, the probability of obtaining y_i for document i is

$$p(y_i) = \sum_{g=1}^{G} \pi_g \, \phi\left(y_i; \boldsymbol{b}_g^T \boldsymbol{\theta}_i, \sigma_g^2\right),$$ [3]

where G is the number of latent classes, π_g is the mixing proportion of latent class g, ϕ denotes the univariate normal density with mean $\boldsymbol{b}_g^T \boldsymbol{\theta}_i$ and variance σ_g^2, and $\boldsymbol{b}_g$ is a vector of regression coefficients on $\boldsymbol{\theta}_i$ for latent class g.

Generative procedure. Given the number of topics (K) and the number of latent classes (G), the generative process of the MixSLDA model is given as

1. For each topic k, draw a $V \times 1$ vector of word distribution $\boldsymbol{\gamma}_k = (\gamma_{k1}, \gamma_{k2}, \dots, \gamma_{kV})^T \sim Dirichlet(\boldsymbol{\beta})$, where V is the number of unique words in the corpus, and $\boldsymbol{\beta}$ is a $V \times 1$ vector of Dirichlet distribution parameters.
2. Draw a $G \times 1$ vector of mixing proportions $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_G) \sim Dirichlet(\boldsymbol{\zeta})$, where $\boldsymbol{\zeta}$ is a $G \times 1$ vector of Dirichlet distribution parameters.
3. For each document i,
    a. Draw a $K \times 1$ vector of topic proportions $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \dots, \theta_{iK}) \sim Dirichlet(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is a $K \times 1$ vector of Dirichlet distribution parameters.
    b. Assign a class membership $C_i = g \sim Multinomial(\boldsymbol{\pi})$.
    c. Draw a label associated with document i, $y_i \sim N\left(\boldsymbol{b}_g^T \boldsymbol{\theta}_i, \sigma_g^2\right)$.

---

[3] Realized values of topic assignments ($\bar{\boldsymbol{z}}_i = \frac{1}{N_i} \sum_{n=1}^{N_i} \tilde{\boldsymbol{z}}_{i,n}$) are used in sLDA to predict the response variable (y_i). Here, we use $\boldsymbol{\theta}_i$, the topic proportions, to calculate information criterion indices DIC and AICM (described in Appendix C) for the MixSLDA for use in determining the number of latent classes using the following set of parameters of interest: $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_D\}$, $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \dots, \boldsymbol{\gamma}_K\}$, $\tilde{\boldsymbol{b}} = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \dots, \boldsymbol{b}_G\}$, $\tilde{\boldsymbol{\sigma}}^2 = \{\sigma_1^2, \sigma_2^2, \dots, \sigma_G^2\}$, and $\boldsymbol{\pi}$.

4. For each word in document i ($w_{i,n}$),

    a. Assign a topic $z_{i,n} \sim Multinomial(\boldsymbol{\theta}_i)$.

    b. Draw a term $w_{i,n} \sim Multinomial(\boldsymbol{\gamma}_{z_{i,n}})$.

Estimation of the MixSLDA. Suppose $\boldsymbol{w} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_D\}$, $\boldsymbol{y} = \{y_1, y_2, \dots, y_D\}$, $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_D\}$, $\boldsymbol{z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \dots, \boldsymbol{z}_D\}$, $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \dots, \boldsymbol{\gamma}_K\}$, $\boldsymbol{C} = \{C_1, C_2, \dots, C_D\}$, $\widetilde{\boldsymbol{b}} = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \dots, \boldsymbol{b}_G\}$, $\widetilde{\boldsymbol{\sigma}}^2 = \{\sigma_1^2, \sigma_2^2, \dots, \sigma_G^2\}$, where D is the number of documents. We use a data augmentation technique (Tanner & Wong, 1987) to easily obtain the posterior distribution of the model parameters. Here, the observed data $\boldsymbol{w}$ and $\boldsymbol{y}$ are augmented with realized values of $\boldsymbol{z}$ and $\boldsymbol{C}$. The joint distribution of $\boldsymbol{w}$, $\boldsymbol{y}$, $\boldsymbol{\theta}$, $\boldsymbol{z}$, $\boldsymbol{\gamma}$, $\boldsymbol{C}$, $\widetilde{\boldsymbol{b}}$, $\widetilde{\boldsymbol{\sigma}}^2$, $\boldsymbol{\pi}$ given the hyper parameters is

$$p\left(\boldsymbol{w}, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{\gamma}, \boldsymbol{C}, \widetilde{\boldsymbol{b}}, \widetilde{\boldsymbol{\sigma}}^2, \boldsymbol{\pi} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\zeta}\right)$$

$$= p(\boldsymbol{\pi}|\boldsymbol{\zeta}) \left[\prod_{i=1}^{D} p\left(y_i | \boldsymbol{\theta}_i, C_i, \widetilde{\boldsymbol{b}}, \widetilde{\boldsymbol{\sigma}}^2\right) p(\boldsymbol{\theta}_i|\boldsymbol{\alpha}) p(C_i|\boldsymbol{\pi}) \left\{ \prod_{n=1}^{N_i} p\left(z_{i,n}|\boldsymbol{\theta}_i\right) p\left(w_{i,n}|z_{i,n}, \boldsymbol{\gamma}\right) \right\} \right]$$

$$\prod_{k=1}^{K} p(\boldsymbol{\gamma}_k|\boldsymbol{\beta}) \left[\prod_{g=1}^{G} p\left(\boldsymbol{b}_g\right) p\left(\sigma_g^2\right) \right]. \qquad (3.2)$$

A Gibbs sampler was used to generate samples from the posterior distribution. Full conditional distributions for implementing the Gibbs sampler are given below. When a full conditional distribution was unknown, a Metropolis-Hastings step was used.

1. The full conditional distribution of $z_{i,n}$ is given as

$$p\left(z_{i,n} = k | \boldsymbol{\theta}_i, w_{i,n} = v, \boldsymbol{\gamma}\right) \propto p\left(z_{i,n} = k | \boldsymbol{\theta}_i\right) p\left(w_{i,n} = v | z_{i,n} = k, \boldsymbol{\gamma}\right) \propto \theta_{ik}\gamma_{kv}.$$

    Thus,

$$\left(z_{i,n} | \boldsymbol{\theta}_i, w_{i,n} = v, \boldsymbol{\gamma}\right) \sim Multinomial(\theta_{i1}\gamma_{1v}, \theta_{i2}\gamma_{2v} \dots, \theta_{iK}\gamma_{Kv}).$$

2. The full conditional distribution of $\boldsymbol{\gamma}_k$ is

$$p(\boldsymbol{\gamma}_k|\boldsymbol{w},\boldsymbol{z},\boldsymbol{\beta}) \propto p(\boldsymbol{\gamma}_k|\boldsymbol{\beta}) \prod_{i=1}^{D} \prod_{n=1}^{N_i} p\left(w_{i,n}|z_{i,n},\boldsymbol{\gamma}_k\right)$$

$$\propto \prod_{v=1}^{V} \gamma_{kv}^{\beta_v - 1} \prod_{v=1}^{V} \gamma_{kv}^{N_{kv}} = \prod_{v=1}^{V} \gamma_{kv}^{N_{kv}+\beta_v-1},$$

where N_{kv} is the number of cases that the word v is assigned to topic k across all documents, and β_v is the vth element of $\boldsymbol{\beta}$. Thus,

$$(\boldsymbol{\gamma}_k|\boldsymbol{w},\boldsymbol{z},\boldsymbol{\beta}) \sim Dirichlet(\boldsymbol{N}_k + \boldsymbol{\beta}),$$

where $\boldsymbol{N}_k = (N_{k1}, \dots, N_{kV})^T$.

3. The full conditional distribution of C_i is

$$p\left(C_i = g|y_i,\widetilde{\boldsymbol{b}},\widetilde{\boldsymbol{\sigma}}^2,\boldsymbol{\pi},\boldsymbol{\theta}_i\right) \propto p(C_i = g|\boldsymbol{\pi}) p\left(y_i|C_i = g,\boldsymbol{b}_g,\sigma_g^2,\boldsymbol{\theta}_i\right)$$

$$\propto \pi_g \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\left(-\frac{1}{2\sigma_g^2}\left(y_i - \boldsymbol{b}_g^T\boldsymbol{\theta}_i\right)^2\right).$$

Thus,

$$\left(C_i\middle|y_i,\widetilde{\boldsymbol{b}},\widetilde{\boldsymbol{\sigma}}^2,\boldsymbol{\pi},\boldsymbol{\theta}_i\right) \sim Multinomial(\tau_1,\tau_2,\dots,\tau_G),$$

where $\tau_g = \pi_g \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\left(-\frac{1}{2\sigma_g^2}\left(y_i - \boldsymbol{b}_g^T\boldsymbol{\theta}_i\right)^2\right)$.

4. The full conditional distribution of $\boldsymbol{\pi}$ is

$$p(\boldsymbol{\pi}|\boldsymbol{C},\boldsymbol{\zeta}) \propto p(\boldsymbol{\pi}|\boldsymbol{\zeta}) p(\boldsymbol{C}|\boldsymbol{\pi})$$

$$\propto \prod_{g=1}^{G} \pi_g^{\zeta_g - 1} \prod_{g=1}^{G} \pi_g^{N_{cg}} = \prod_{g=1}^{G} \pi_g^{\zeta_g + N_{cg} - 1},$$

where ζ_g is the gth element of $\boldsymbol{\zeta}$, and N_{cg} is the number of documents that were assigned to latent class g. Thus,

$$(\boldsymbol{\pi}|\boldsymbol{C},\boldsymbol{\zeta}) \sim Dirichlet(\boldsymbol{N}_{cls} + \boldsymbol{\zeta}),$$

where $\boldsymbol{N}_{cls} = (N_{c1}, \dots, N_{cG})$.

5. Suppose $\boldsymbol{b}_g \sim N(\boldsymbol{b}_0, \boldsymbol{\Sigma}_0)$ as a prior. The full conditional distribution of $\boldsymbol{b}_g$ is

$$p(\boldsymbol{b}_g|\boldsymbol{y}, \sigma_g^2, \boldsymbol{\theta}, \boldsymbol{C}) \propto p(\boldsymbol{b}_g) \prod_{i \in D_g} p(y_i|\boldsymbol{\theta}_i, C_i, \boldsymbol{b}_g, \sigma_g^2)$$

$$\propto \exp\left\{-\frac{1}{2}(\boldsymbol{b}_g - \boldsymbol{b}_0)^T \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{b}_g - \boldsymbol{b}_0)\right\} \prod_{i \in D_g} \exp\left\{-\frac{1}{2\sigma_g^2}(y_i - \boldsymbol{b}_g^T \boldsymbol{\theta}_i)^2\right\}.$$

Thus,

$$(\boldsymbol{b}_g|\boldsymbol{y}, \sigma_g^2, \boldsymbol{\theta}, \boldsymbol{C}) \sim N\left(\boldsymbol{\mu}_{b_g}, \boldsymbol{V}_{b_g}\right),$$

where $\boldsymbol{V}_{b_g} = \left(\boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma_g^2}\sum_{i \in D_g} \boldsymbol{\theta}_i \boldsymbol{\theta}_i^T\right)^{-1}$ and $\boldsymbol{\mu}_{b_g} = \boldsymbol{V}_{b_g}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{b}_0 + \frac{1}{\sigma_g^2}\sum_{i \in D_g} \boldsymbol{\theta}_i y_i\right)$, in which D_g is a set of documents that were assigned to class g.

6. Suppose $\sigma_g^2 \sim InvGamma(c_0, d_0)$ as a prior. The full conditional distribution of σ_g^2 is

$$p(\sigma_g^2|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{b}_g, \boldsymbol{C}) \propto p(\sigma_g^2) \prod_{i \in D_g} p(y_i|\boldsymbol{\theta}_i, C_i, \boldsymbol{b}_g, \sigma_g^2)$$

$$\propto (\sigma_g^2)^{-c_0 - 1} \exp\left(-\frac{d_0}{\sigma_g^2}\right) \prod_{i \in D_g} \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\left\{-\frac{1}{2\sigma_g^2}(y_i - \boldsymbol{b}_g^T \boldsymbol{\theta}_i)^2\right\}.$$

Thus,

$$(\sigma_g^2|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{b}_g, \boldsymbol{C}) \sim InvGamma\left(\frac{N_{cg}}{2} + c_0, d_0 + \frac{1}{2}\sum_{i \in D_g}(y_i - \boldsymbol{b}_g^T \boldsymbol{\theta}_i)^2\right).$$

7. The full conditional distribution of $\boldsymbol{\theta}_i$ is

$$p(\boldsymbol{\theta}_i|\boldsymbol{z}, \widetilde{\boldsymbol{b}}, \widetilde{\boldsymbol{\sigma}}^2, C, y_i, \boldsymbol{\alpha}) \propto p(\boldsymbol{\theta}_i|\boldsymbol{\alpha}) p(y_i|\boldsymbol{\theta}_i, \boldsymbol{b}_g, \sigma_g^2, C_i = g) \prod_{n=1}^{N_i} p(z_{i,n}|\boldsymbol{\theta}_i)$$

$$\propto \prod_{k=1}^{K} \theta_{ik}^{\alpha_k - 1} \prod_{k=1}^{K} \theta_{ik}^{N_{ik}} \exp\left(-\frac{1}{2\sigma_g^2}(y_i - \boldsymbol{b}_g^T \boldsymbol{\theta}_i)^2\right)$$

$$\propto \prod_{k=1}^{K} \theta_{ik}^{N_{ik}+\alpha_k-1} \exp\left(-\frac{1}{2\sigma_g^2}\left(-2y_i \boldsymbol{b}_g^T \boldsymbol{\theta}_i + \left(\boldsymbol{b}_g^T \boldsymbol{\theta}_i\right)^2\right)\right), \qquad (3.3)$$

where N_{ik} is the number of words that are assigned to topic k in document i. Since the full conditional distribution of $\boldsymbol{\theta}_i$ is unknown, a Metropolis-Hastings algorithm was used to sample $\boldsymbol{\theta}_i$ as follows:

After the mth iteration,

1) Generate a candidate $\boldsymbol{\theta}_i^{(new)} \sim Dirichlet\left(\psi \boldsymbol{\theta}_i^{(m)}\right)$, where ψ is a tuning parameter for adjusting the acceptance rate of candidates and producing MCMC samples with good mixing properties. Recommended acceptance rate is about 0.44 for unidimensional sampling and decreases to 0.23 for high dimensions (Gelman, Roberts, & Gilks, 1996).
2) Take

$$\boldsymbol{\theta}_i^{(m+1)} = \begin{cases} \boldsymbol{\theta}_i^{(new)} \text{ with probability } \alpha = \min\left\{\dfrac{f\left(\boldsymbol{\theta}_i^{(new)}\right) g\left(\boldsymbol{\theta}_i^{(m)} | \boldsymbol{\theta}_i^{(new)}\right)}{f\left(\boldsymbol{\theta}_i^{(m)}\right) g\left(\boldsymbol{\theta}_i^{(new)} | \boldsymbol{\theta}_i^{(m)}\right)}, 1\right\} \text{ or} \\ \boldsymbol{\theta}_i^{(m)} \text{ otherwise} \end{cases},$$

where $g(\cdot\,|\boldsymbol{\theta}_i) = Dirichlet(\boldsymbol{\theta}_i)$, and

$$f(\boldsymbol{\theta}_i) = \prod_{k=1}^{K} \theta_{ik}^{N_{ik}+\alpha_k-1} \exp\left(-\frac{1}{2\sigma_g^2}\left(-2y_i \boldsymbol{b}_g^T \boldsymbol{\theta}_i + \left(\boldsymbol{b}_g^T \boldsymbol{\theta}_i\right)^2\right)\right).$$

Illustrative Example

In this section, we illustrate the use of the MixSLDA with data from an NSF-funded project designed to train teachers to teach science inquiry skills to middle grades and high school students. In this example, we address the following research questions using a dataset obtained

from a science inquiry test that was administered as a part of the project: (i) What are the latent themes underlying students' answers? (ii) Are there any subgroups that have different relationships between topic proportions and scores? (iii) What are the relationships between topic proportions and scores?

Test

The test was designed to measure middle grades students' use of academic language and understanding of science inquiry practices on physical science experiments. The test specifically focused on the use of cause and effect, independent and dependent variables, and construction of hypotheses in students' reasoning. The test consisted of seven constructed response testlets, each with from two to five items. Each item was scored with from two to four points. The total of the item scores over all seven testlets was used as the score variable for the MixSLDA.

Tests were administered once at the beginning of the year as a pre-test and a second time at the end of the year as a post-test. To counter memory effects between pre-test and post-test, two different forms of the test were constructed, Form A and Form B. Both forms had parallel testlet structures. One of the two forms was randomly given to students for the pre-test, and the alternate form was given for the post-test. For purposes of this example, pre-test and post-test responses from only Form A were used in this analysis. Each student's answers to the constructed response items from all seven testlets were combined and considered as one answer document. Thus, in this analysis, each student had one answer document with an associated score. Note that students in the pre-test data and students in the post-test data were different as students alternated forms between the pre- and post-tests.

Sample

The sample for this example consisted of 530 documents: 276 documents (52.08%) were from the pre-test, and 254 documents (47.92%) were from the post-test. There were 277 answer documents (52.26%) from 7th grade students, 252 answer documents (47.55%) from 8th grade students, and one document (0.19%) from a 6th grade student. Of these 530 documents, 243 documents (45.85%) were produced by female students, and 287 documents (54.15%) by male students. The mean score for the 530 samples was 31.87 with a standard deviation of 14.54. A density plot of the scores presented in Figure 3.1 shows that the score distribution has more than one mode, which suggests that the use of a mixture distribution might better describe the data.

Pre-Processing of Data

Before starting the analysis, the text data from the 530 students were preprocessed. First, stopwords in the dataset were removed. Stopwords are words which tend to appear with high frequency, but which do not have significant meaning, such as "a," "the," "and," etc. Then, all remaining words were converted to their basic roots. For example, plural terms were converted to singular form, capitalized words to small letters, and typos corrected. After the preprocessing, the corpus of 530 documents consisted of 415 unique words. The average document length was approximately 86 words with a standard deviation of about 37 words.

Determining the Number of Topics in the Model

LDA was used to determine the number of latent topics in students' written responses. LDA models with from two to seven topics were fitted to the data using a Gibbs sampling algorithm (described in Appendix C). The maximum number of topics was set to seven because the content of the test was limited by seven science scenarios.

Two information indices were used to inform model selection: DIC (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) and AICM (Raftery, Newton, Satagopan, & Krivitsky, 2007). These were calculated for each of the six LDA models. (The equations for DIC and AICM are given in Appendix C.) DIC and AICM values are presented in Table 3.1. Smaller DIC values and larger AICM values suggest better model fit relative to the candidate models. The seven topic model had the smallest DIC value, and the four topic model had the largest AICM value. In addition, words with high mean posterior probabilities were examined for all topics in each model. Models with more than four topics had some topics that appeared to be redundant, that is, the high frequency words in these models were similar to one another. Therefore, $K = 4$ was taken as the more usefully interpretable number of topics given the students' responses.

Assuming No Latent Classes

The MixSLDA with no latent classes is equivalent to sLDA. Assuming there were four topics ($K = 4$) in the corpus, the following parameters were used to construct the Gibbs sampler: $\boldsymbol{\beta} = \boldsymbol{j}_V$, $\boldsymbol{\alpha} = \boldsymbol{j}_K$, $\boldsymbol{b}_0 = \mathbf{0}$, $\boldsymbol{\Sigma}_0 = 100^2\boldsymbol{I}_K$, $c_0 = 3$, and $d_0 = 60$. The value of ψ was tuned to have an acceptance rate of approximately 0.4 for the Metropolis-Hastings step. In total, 40,000 iterations were run with the first 20,000 iterations being discarded as burn-in, and the next 20,000 iterations used for estimating the posterior. Figures 3.2 and 3.3 show the trace plots of $\boldsymbol{b}$ and σ^2, respectively. The trace plots appeared to be stabilized and mixed well after the burning period. Convergence was checked by Geweke (Geweke, 1992) and Heidelberger and Welch (Heidelberger & Welch, 1983) convergence diagnostics using the R package *coda* (Plummer, Best, Cowles, & Vines, 2006).

Table 3.2 presents topics obtained from this analysis. Words in the Topic 1 column were mostly about questions for "controlling variables," and words in the Topic 2 column were

generally about questions for "hypothesis, observation, and evidence." Words in the Topic 3 column were mostly about questions for "cause and effect," and words in the Topic 4 column were mostly related to everyday language. Table 3.3 shows the average of the posterior mean estimates of $\boldsymbol{\theta}_i$ and the posterior mean estimates of regression coefficient parameters ($\boldsymbol{b}$). The posterior mean estimate of σ^2 was 58.89. The average values of four topic proportion estimates ($\boldsymbol{\theta}$ column in Table 3.3) were quite similar. The posterior mean estimate of Topic 2 regression coefficient was the highest among the four components of $\boldsymbol{b}$, which implied that Topic 2 was highly related to test score. The Topic 2 regression coefficient estimate 70.72 indicates that if a student's responses were from Topic 2, the expected score would be 70.72. For Topics 1, 3, and 4, the expected scores were 35.09, 14.04, and 9.88, respectively.

Assuming Two Latent Classes

The following parameters were used to construct the Gibbs sampler for the MixSLDA model with four topics and two latent classes: $\boldsymbol{\beta} = \boldsymbol{j}_V$, $\boldsymbol{\alpha} = \boldsymbol{j}_K$, $\boldsymbol{b}_0 = \mathbf{0}$, $\Sigma_0 = 100^2\boldsymbol{I}_K$, $c_0 = 0.001$, $d_0 = 0.001$, and $\boldsymbol{\zeta} = 5\boldsymbol{j}_2$. As with the previous analysis, the value of ψ was set to have an average acceptance rate of 0.4 in the Metropolis-Hastings step. In total, 40,000 iterations were run with the first 20,000 iterations being discarded as burn-in, and the next 20,000 iterations were used for posterior inferences. Figures 3.4, 3.5, and 3.6 show the trace plots of $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, and σ_1^2, and σ_2^2, respectively. All the trace plots appeared to be stabilized and showed good mixing after the burn-in period.

Table 3.4 presents the four topics obtained from this analysis. Inspection of the words in each topic demonstrated that the topics obtained from the sLDA and the topics obtained from the MixSLDA with two latent classes were quite similar. For each document, a posterior mode of latent class membership was obtained to investigate the characteristics of two latent classes: 309

documents were assigned to Class 1, and 221 documents were assigned to Class 2. The mean of scores for documents in Class 1 and Class 2 were 26.65 and 39.16, respectively, indicating that documents in Class 2 tended to be associated with higher scores than documents in Class 1.

Table 3.5 shows the average values of the posterior mean estimates of the topic proportions, $\boldsymbol{\theta}_i$, and the posterior mean estimates of the regression coefficients ($\boldsymbol{b}_1$ and $\boldsymbol{b}_2$) for each of the two latent classes. The average topic proportions suggest that students in the two classes tended to use different proportions of topics to construct their answers. For example, Topic 4 had the smallest average proportion in Class 2, but had the largest average proportion in Class 1. The posterior mean estimates of regression coefficients indicated that Class 1 and Class 2 had different relationships between topic proportions and scores. The regression coefficient estimate for Topic 1 was 21.48 for Class 1 and 54.51 for Class 2, meaning that an expected score by answering items using words from Topic 1 was 21.48 for students in Class 1 and 54.51 for students in Class 2. Regression coefficient estimate for Topic 2 was 76.71 for Class 1 and 73.06 for Class 2, which indicated that both classes appeared to have high expected scores by using words from Topic 2. The effects of Topic 3 and Topic 4 on the score variable were weak for both classes as can be seen in the regression coefficient estimates for Topics 3 and 4. The posterior mean estimates of σ_1^2 and σ_2^2 were 28.48 and 33.87 for Class 1 and Class 2, respectively.

Determining the Number of Latent Classes

The MixSLDA model assumes that the number of latent classes is known. The previous two analyses illustrated a MixSLDA analysis assuming no latent classes and MixSLDA analysis assuming two latent classes. In order to decide the reasonable number of latent classes for the student response data, DIC and AICM from MixSLDA models with no latent class to three latent classes were calculated to inform model selection. (Details on calculation of DIC and AICM for

the MixSLDA are given in Appendix C.) As can be seen in Table 3.6, DIC and AICM had the smallest values for the two-latent-class model. Further, since the density plot of scores shown in Figure 3.1 appears to have two modes, the two-latent-class model seems reasonable.

Simulation Study

In this section, results of a simulation study are reported for the MixSLDA under testing conditions reflecting the features of the previous empirical example. In this study, we investigated whether the (i) topic structure, (ii) document sizes, and (iii) document lengths had an impact on the MixSLDA parameter recovery.

Simulation Study Design

Following the empirical example, in this simulation study, the number of topics was set to 4 ($K = 4$), and the number of unique words was set to 400 ($V = 400$). The number of latent classes was fixed at 2 ($G = 2$), and the true values of regression coefficients for two latent classes were $\boldsymbol{b}_1 = (20, 70, 15, 20)^T$ and $\boldsymbol{b}_2 = (55, 70, 35, 20)^T$, similar to those obtained from the empirical data analysis. Mixing proportions for the two latent classes were equally assigned ($\boldsymbol{\pi} = (0.5, 0.5)^T$). Topic proportions for documents in each latent class were generated from a $Dirichlet(\boldsymbol{\alpha})$ with $\boldsymbol{\alpha} = (0.25, 0.25, 0.25, 0.25)^T$. In the empirical data analysis, estimated topic proportions were all larger than 0, mostly larger than 0.05. Therefore, topic proportions for this simulation study were generated to have values larger than 0.05 to resemble the student response data.

Word probabilities for the four topics were generated by considering topic similarities. This information was measured using the following formula (Deveaud, SanJuan, & Bellot, 2012; Hong & Davison, 2010; Steyvers & Griffiths, 2007; Yan, 2014):

$$\bar{JS} = \frac{1}{K(K-1)} \sum_{k,k' \in \{1,\ldots,K\}} D(\gamma_k || \gamma_{k'}),$$

where k and k' are distinct values, and $D(\gamma_k||\gamma_{k'})$ is a Jensen-Shannon divergence measure (Lin, 1991), which measures similarities between topics (e.g., γ_k and $\gamma_{k'}$). $D(\gamma_k||\gamma_{k'})$ is defined as:

$$D(\gamma_k||\gamma_{k'}) = 0.5 * \sum_v p(v|k) \log \frac{p(v|k)}{m} + 0.5 * \sum_v p(v|k') \log \frac{p(v|k')}{m},$$

where $m = 0.5 * (p(v|k) + p(v|k'))$, and $p(v|k)$ is the probability of word v appearing in a document under topic k, which is expressed as $p(v|k) = \gamma_{kv}$. Two different sets of topics were generated based on this measure. The first set of topics (labeled as condition S1) were generated to have a $\overline{JS}$ value that was similar to the $\overline{JS}$ value of topics obtained from the empirical data analysis, which was approximately 0.20. The second set of topics (labeled as condition S2) were generated to be more distinct than the first set of topics, and had a $\overline{JS}$ value of 0.35. For both sets of topics, Topic 1 was generated to have high probability on the first 100 words and low probability on the rest, Topic 2 was generated to have high probability on the second 100 words and low probability on the rest, Topics 3 was generated to have high probabilities on the third 100 words, and Topic 4 was generated in a similar manner for the next 100 words.

In order to investigate the effect of document sizes and document lengths on parameter recovery, three document sizes ($D = 300, 500$, and 1000 words), and three average document lengths ($L = 50, 100$, and 200 words) were simulated. In each combination of these topic structures, document sizes, and document lengths, scores were generated from a mixture of two univariate normal distributions. The variances of the normal distributions (i.e., residual variances) were constrained to be the same across the two latent classes for simplicity ($\sigma^2 = \sigma_1^2 = \sigma_2^2$), and two levels of class separation were set by varying the variances. Class separation was measured by the Mahalanobis distance, which is given as $\Delta = |\mu_1 - \mu_2|/\sigma$, where μ_1 and μ_2

are the means of the two mixture components. The μ_1 was calculated using $\boldsymbol{b}_1^T\overline{\boldsymbol{\theta}}_1$, where $\overline{\boldsymbol{\theta}}_1$ is a vector of topic proportions that were averaged across documents belonging to Class 1. Similarly, the μ_2 was calculated using $\boldsymbol{b}_2^T\overline{\boldsymbol{\theta}}_2$. Two levels of Δ were considered: 2.5 and 3.5, which corresponds to $\sigma^2 = 30$ and $\sigma^2 = 15$, respectively. In total, there were 36 conditions simulated (= 2 topic structures × 3 number of documents × 3 document lengths × 2 class separations) in the study.

Estimation

The number of topics and the number of latent classes were assumed to be known. For each condition, the Gibbs sampler was used with the following prior distributions:

1) $\boldsymbol{\gamma}_k \sim Dirichlet(\boldsymbol{\beta})$ with $\boldsymbol{\beta} = \boldsymbol{j}_V$ for $k = 1,2,3,4$.
2) $\boldsymbol{\pi} \sim Dirichlet(\boldsymbol{\zeta})$ with $\boldsymbol{\zeta} = 10\boldsymbol{j}_2$.
3) $\boldsymbol{b}_g \sim N(\boldsymbol{b}_0, \boldsymbol{\Sigma}_0)$ with $\boldsymbol{b}_0 = (0,0,0,0)^T$ and $\boldsymbol{\Sigma}_0 = 100^2 * I_K$.
4) $\sigma^2 \sim InvGamma(c_0, d_0)$ with $c_0 = 3$ and $d_0 = 60$ when data were generated with $\sigma^2 = 30$ and with $c_0 = 3$ and $d_0 = 30$, when data were generated with $\sigma^2 = 15$.
5) $\boldsymbol{\theta}_i \sim Dirichlet(\boldsymbol{\alpha})$ with $\boldsymbol{\alpha} = \boldsymbol{j}_K$.

The total number of iterations was 40,000 with the first 15,000 iterations used as burn-in. The subsequent 25,000 iterations were used for estimating the posterior. For those cases in which data were generated with $\Delta = 2.5$, $\overline{JS} = 0.20$, and $L = 50$, the total number of iterations was 50,000 with 15,000 iterations as burn-in. Parameter estimates were summarized using the 100 converged replications. Convergence was evaluated by Geweke (Geweke, 1992), Heidelberger and Welch (Heidelberger & Welch, 1983) diagnostics and also by examination of trace plots. The average acceptance rate for sampling $\boldsymbol{\theta}_i$ in the Metropolis-Hastings step was approximately 0.4.

In the course of estimation, label switching can arise because the likelihood is invariant under the permutation of labels $g = 1, \ldots, G$ (McLachlan & Peel, 2000). When label switching occurs, the switched labels need to be renamed for posterior inferences. This can be handled by restricting the parameter space, such as a constraint on mixing proportion components or a constraint on mean components (Gelman, Carlin, Stern, Dunson, Vehtari, & Rubin, 2013; McLachlan & Peel, 2000). In this analysis, in order to avoid the possibility of label switching, a restriction was imposed on the regression coefficients corresponding to Topic 1 such that b_{11} was constrained to be less than b_{21}.

Simulation Study Results

For each of the simulation study conditions, posterior mean estimates of regression coefficients and the residual variance were summarized by examining the mean absolute bias, mean relative bias, standard error, and coverage rate of the 95% credibility intervals. The general form of absolute bias (AB) is defined as $AB = |\omega - \hat{\omega}|$, where $\hat{\omega}$ is a posterior mean estimate of ω. The general form of relative bias (RB) is $RB = 100 * \left(\frac{\hat{\omega}-\omega}{\omega}\right)$, which is a percentage of bias relative to the true parameter value. Following previous research (e.g., Curran, West, & Finch, 1996; Finch, West, & MacKinnon, 1997; Flora & Curran, 2004; Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009; Kaplan, 1989), RB values less than 5% were considered as trivial bias, RB values between 5% and 10% were considered as moderate bias, and RB values larger than 10% were considered as substantial bias. The coverage rate was defined as $CP = \frac{1}{R} Ind\left(l_{\hat{\omega}_{(r)}} \leq \omega \leq u_{\hat{\omega}_{(r)}}\right)$, where R is the number of replications, $l_{\hat{\omega}_{(r)}}$ and $u_{\hat{\omega}_{(r)}}$ are the lower and upper bound of the 95% credibility interval for $\hat{\omega}_{(r)}$, respectively, $\hat{\omega}_{(r)}$ is a posterior mean estimate of ω obtained from the rth replication, and $Ind(\cdot)$ is an indicator function scored as 1 when ω was

within the interval and 0 otherwise. For each of $\boldsymbol{b}_1$ components, $\boldsymbol{b}_2$ components, and σ^2, the mean AB, the mean RB, coverage rate, and standard error across 100 replications were calculated for each condition.

Tables 3.7 and 3.8 summarize mean AB and mean RB values for the posterior mean estimates of $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ components under the S1 topic condition. In general, the mean AB and mean RB values of $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ components decreased as the document size or document length increased for both mode separations. For each condition, the mean RB for b_{13} generally had the largest value among the four $\boldsymbol{b}_1$ regression components (except for the condition with $\Delta = 2.5$, $D = 300$, and $L = 50$, and the condition with $\Delta = 3.5$, $D = 500$, and $L = 50$), and the mean RB for b_{24} had the largest value among the four $\boldsymbol{b}_2$ components. Further investigating the mean AB values for $\boldsymbol{b}_1$ indicated that the absolute bias values for b_{11}, b_{13}, and b_{14} were generally similar, but the generating value for b_{13} was the smallest among the four $\boldsymbol{b}_1$ components ($\boldsymbol{b}_1 = (20, 70, 15, 20)^T$), resulting in larger relative bias than the other components. In the case of $\boldsymbol{b}_2$, b_{24} had the largest mean RB due to the same reason. The mean RB for $\boldsymbol{b}_1$ components were less than approximately 10% when the average document lengths was larger than 50 (i.e., $L = 100$ and $L = 200$), and the mean RB for $\boldsymbol{b}_2$ components had less than approximately 5% except for b_{24} when the average document length was larger than 50 (b_{24} ranged from 5.11% to 18.07%).

Tables 3.9 and 3.10 summarize the standard errors and coverage rates for the posterior mean estimates of $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ components under the S1 topic condition. For both mode separations, standard errors tended to decrease as the document size or document length increased. In the case of coverage rate, most of the coverage rates for b_{11}, b_{13}, b_{14}, b_{21}, b_{22}, b_{23} were larger than about 0.90. The coverage rates for b_{12} and b_{24} ranged from 0.54 to 0.90 and 0.67 to 0.94, respectively. These coverage rates increased as the average document length

increased within a given document size, and decreased as the document size increased within a given document length, especially when $L = 50$. Inspection of the trace plots of b_{12} and b_{24} under the $L = 50$ condition indicated that the widths of 95% credibility intervals for the two parameter estimates decreased markedly compared to the reduction in the size of bias, resulting in smaller coverage rates as document size increased.

Results for σ^2 and latent class membership recovery under the S1 topic condition are presented in Table 3.11. The mean AB, mean RB, and standard error values for the posterior mean estimate of σ^2 appeared to decrease as the document size or document length increased. The mean RB values of σ^2 estimate were less than 10% when the average document length was larger than 50 for the $\Delta = 2.5$ condition. On the other hand, the mean RB values were less than 10% when the average document length was larger than 100 for the $\Delta = 3.5$ condition. Overall, the coverage rate increased as the document length increased, and decreased as the document size increased, especially when $L = 50$ due to the same reason as discussed previously for b_{12} and b_{24}. With respect to the latent class membership recovery, the percentage correct increased as document size or document length increased. As might be expected, the percentages for $\Delta = 3.5$ were larger than those for $\Delta = 2.5$.

Tables 3.12 and 3.13 present the mean AB and the mean RB for the posterior mean estimates of $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ components under the S2 topic condition. Within each of the three document sizes, the mean AB appeared to be similar across three document length conditions. For a given average document length, the mean AB decreased with increasing document size. As for the mean RB, the average document length did not have an apparent effect on the mean RB values for $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ components, but the mean RB tended to decrease as the document size increased. The mean RB values of b_{11} and b_{12} were less than 5%, and b_{13} and b_{14} had mean RB

values less than approximately 10%. The mean RB values for b_{21}, b_{22}, b_{23} were all less than 5%, and the mean RB values for b_{24} were less than approximately 10%.

Tables 3.14 and 3.15 summarize the standard errors and coverage rates for the posterior mean estimates of $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ components under the S2 topic condition. Standard errors appeared to decrease as the document size or document lengths increased for both mode separations. Most of the coverage rates were larger than 0.90. Coverage rates under 0.90 were mostly from the conditions with $L = 200$.

Table 3.16 presents the results for σ^2 and latent class membership recovery under the S2 topic condition. The mean AB and standard errors tended to decrease as the document size or document length increased. The mean RB also tended to decrease as the average document lengths increased for a given document size. The mean RB appeared to be influenced by document size, only when the average document length was 50. When the average document length was larger than 50, the document size did not seem to have an apparent effect on the mean RB. In case of coverage rate, it increased as the average document length increased for a given document size. The document size did not appear to have any clear effects on the coverage rate. As for the latent class membership recovery, it increased as the document size increased or as the document length increased. The latent class membership recovery values for $\Delta = 3.5$ were larger than those for $\Delta = 2.5$.

Simulation Study Conclusions

A simulation study was conducted to investigate whether the MixSLDA parameter recovery was affected by topic structure, document size, and document length. The results of this simulation study provided some useful suggestions regarding the three questions addressed at the beginning of this study. In order to summarize results from this study more efficiently, for each

mean AB, mean RB, standard errors, and coverage rates for $\boldsymbol{b}_1 = (b_{11}, b_{12}, b_{13}, b_{14})$, the values of its four components were individually averaged; the same averaging was performed for $\boldsymbol{b}_2$. The averaged values for $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are presented in Tables 3.17 and 3.18, respectively.

First, topic structures appeared to influence the estimation of regression parameters. The S1 topics were more similar to one another than were the S2 topics to one another. In general, the mean AB, mean RB, and standard errors under the S2 topic condition were smaller than those under the S1 topic condition, particularly when the average document length was small (e.g., $L = 50$). In addition, the coverage rates for the S2 condition tended to be higher than the coverage rates for the S1 condition.

Second, the mean AB, mean RB, and standard errors appeared to be affected by document size and average document length. In general, these statistics decreased as document size or document length increased, suggesting the total number of words in a corpus is important for obtaining accurate parameter estimates. For the S1 topic structure and $\Delta = 2.5$ (see Table 3.17), the average of mean RB values of $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ components (here we designate this as Mean RB) were larger than 10%, when $D = 300$ & $L = 50$ and $D = 500$ & $L = 50$. This suggests that these particular combinations of document size and document length may not be sufficient in order to obtain accurate parameter estimates. For the S2 topic structure (see Table 3.18), the Mean RB values for $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ were all less than about 5%. Both Mean AB and Mean RB values decreased as document size increased, but the values did not seemed to be influenced by document length. This may suggest that increasing document length greater than 50 does not necessarily result in $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ estimates with less bias when topics are similar to the S2 topics. The mean RB values for variance estimates were larger than 10% when $L = 50$ across all document sizes, topic structures, and mode separations.

The mean RB values for b_{13} tended to be higher than the mean RB values for the other regression coefficients in the corresponding latent class; the same was true for b_{24}. These two coefficients had the smallest true values of regression coefficients within the corresponding latent class. For the S2 topic structure, the mean RB values for these two parameters were mostly less than 10% across all document sizes, document lengths, and mode separations. For the S1 topic structure, the mean RB values for b_{13} and b_{24} showed substantial bias when $L = 50$, that is, the mean RB for b_{13} ranged from 2.94% to 17.62% and the mean RB for b_{24} ranged from 19.39% to 33.13%. This suggested that the estimation of the smallest regression coefficient might not be accurate when the average document length is not larger than 50.

The percentage of correct latent class memberships were greater than 80% except for the conditions with $\Delta = 2.5, D = 300, L = 50$ and $\Delta = 2.5, D = 500, L = 50$ under the S1 topic condition. The percentages increased as the document size or document length increased. As might be expected, the percentages for $\Delta = 3.5$ were greater than those for $\Delta = 2.5$, although the differences were not large.

Discussion

In this study, a finite mixture supervised latent Dirichlet allocation model, a MixSLDA model, was introduced for analyzing text data and associated scores from constructed response items. This model extended the supervised latent Dirichlet allocation model by allowing the score variable to be represented by a finite set of regression equations, with each equation corresponding to a unique relationship between topic proportions and scores.

The MixSLDA model offers the possibility of providing a rapid way to analyze the relationship between students' scores on their constructed response items and their writing. In addition, because of the complexity of writing, the model permits analyzing students'

performance on constructed response items and tests in order to account for the heterogeneity present in the population. The MixSLDA does this by detecting latent groups of examinees (or documents) that have different relationships between topic proportions and a score (or label) variable. These latent groups are homogeneous on the variable(s) that caused the groups to form.

The performance of the MixSLDA was illustrated using an empirical dataset. The model provided evidence that the students' documents could be summarized by two latent classes that differed in their relationship between the scores and the four latent topics detected in the dataset. The main difference between the two latent classes was in the relationship between the use of Topic 1 and the score variable: the use of Topic 1 led to a higher expected score for Class 2, but not for Class 1. This result suggests that documents in Class 1 with high Topic 1 proportions failed to adequately discuss the problems posed in Topic 1 related questions.

Results of a simulation study conducted to evaluate the MixSLDA model performance suggested that having either a large number of documents or having documents with a relatively large number of words is necessary in order to obtain accurate parameter estimates. In other words, the total number of words in a corpus was important for parameter estimation. This was especially the case when topics in the data were similar to one another (e.g., the S1 topic structure). Educational assessments often have constructed response items that do not require a long response. Results of this simulation study suggested that inaccuracy in estimating the MixSLDA due to short answers may be overcome by collecting larger numbers of answer documents.

The latent classes detected using the MixSLDA are dependent in part on the functions used for describing the relationship between topic proportions and scores (Gelman et al., 2013). Given this caveat, the MixSLDA model can be useful for understanding and summarizing the

relationship between examinees' writing and scores in that it provides a flexible approach that helps to account for heterogeneity in test performances. In addition, this study was conducted with a continuous score variable. Extensions of MixSLDA to include a categorical type of score variable, such as by use of a generalized linear model (Blei & McAuliffe, 2008) for sLDA, would be useful.

References

Andrade, H. L., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education: Principles, Policy & Practice*, *17*(2), 199-214.

Bisgin, H., Liu, Z., Fang, H., Xu, X., & Tong, W. (2011). Mining FDA drug labels using an unsupervised learning technique-topic modeling. In *BMC bioinformatics* (Vol. 12, No. 10, p. S11). BioMed Central.

Blei, D.M. (2012). Probabilistic topic models. Communications of the ACM, 55, 77-84.

Blei, D. M., & McAuliffe, J. D. (2008). Supervised topic models. In *Advances in neural information processing systems* (pp. 121-128).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*, 993-1022.

Chen, Y., Yu, B., Zhang, X., & Yu, Y. (2016). Topic modeling for evaluating students' reflective writing: a case study of pre-service teachers' journals. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 1-5). ACM. Retrieved from https://doi.org/10.1145/2883851.2883951

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological methods*, *1*(1), 16-29.

Deveaud, R., SanJuan, E., & Bellot, P. (2012). LIA at TREC 2012 Web track: Unsupervised search concepts identification from general sources of information: *Proceedings of the 21th Text Retrieval Evaluation Conference*, Gaithersburg, MD. Retrieved from https://apps.dtic.mil/dtic/tr/fulltext/u2/a581250.pdf

Finch, J. F., West, S. G., & MacKinnon, D. P. (1997). Effects of sample size and nonnormality on the estimation of mediated effects in latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal*, *4*(2), 87-107.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological methods*, *9*(4), 466-491.

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, *16*(4), 625-641.

Furtak, E. M., & Ruiz-Primo, M. A. (2008). Making students' thinking explicit in writing and discussion: An analysis of formative assessment prompts. *Science Education*, *92*(5), 799-824.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman and Hall/CRC.

Gelman, A., Roberts, G. O., & Gilks, W. R. (1996). Efficient Metropolis jumping rules. *Bayesian statistics, 5* (pp. 599-608). Oxford: Oxford University Press.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (with discussion). In J. O., Berger, J. M. Bernardo, A. P. Dawid, & A. F. M. Smith (Eds.) *Bayesian Statistics 4* (pp. 156-163). Oxford: Oxford University Press.

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management, 59*, 467-483.

Heidelberger, P., & Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, *31*(6), 1109-1144.

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics* (pp. 80-88). ACM.

Kaplan, D. (1989). A study of the sampling variability and z-values of parameter estimates from misspecified structural equation models. *Multivariate Behavioral Research*, *24*(1), 41-57.

Kim, S., Kwak, M., Cardozo-Gaibisso, L., Buxton, C., & Cohen, A. S. (2017). Statistical and Qualitative Analyses of Students' Answers to a Constructed Response Test of Science Inquiry Knowledge. *Journal of Writing Analytics*, *1*, 82-102.

Lacoste-Julien, S., Sha, F., & Jordan, M. I. (2008). DiscLDA: discriminative learning for dimensionality reduction and classification. In *Proceedings of the 21st International Conference on Neural Information Processing Systems* (pp. 897-904). Curran Associates Inc.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, *37*(1), 145-151.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Mimno, D., & McCallum, A. (2008). Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Proceedings of the 24th conference on Uncertainty in artificial intelligence* (pp. 411-418). AUAI Press.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R news*, *6*(1), 7-11.

Prier, K. W., Smith, M. S., Giraud-Carrier, C., & Hanson, C. L. (2011). Identifying health-related topics on Twitter. In J. Salerno, S. J. Yang, D. Nau, & SK. Chai (Eds.), *Social*

Computing, Behavioral-Cultural Modeling, and Prediction (pp. 18-25). Springer, Berlin, Heidelberg.

Raftery, A. E., Newton, M. A., Satagopan, J. M., & Krivitsky, P. N. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity (with Discussion) In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, & M. West (Eds.) *Bayesian Statistics 8*. Oxford: Oxford University Press.

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, *111*(515), 988-1003.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence* (pp. 487-494). AUAI Press.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583-639.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, *427*(7), 424-440.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, *82*(398), 528-540.

Yan, E. (2014). Research dynamics: Measuring the continuity and popularity of research topics. *Journal of Informetrics*, *8*(1), 98-110.

CHAPTER 4

TOPIC MODELING FOR LONGITUDINAL TEXT DATA[4]

[4] Kim, S., Lu, Z., and Cohen, A.S. To be submitted to *Multivariate Behavioral Research*

Abstract

In the social and behavioral sciences, text data are available from numerous sources, such as standardized achievement tests and research projects. This is because constructed response items are becoming increasingly prominent as a way of enabling examinees to show their reasoning. In order to analyze text data, topic modeling has been widely used across disciplines. Topic modeling is a natural language statistical technique for investigating the latent thematic structure that underlies the text in a collection of documents. This study propose the growth curve topic model for analyzing examinees' written responses collected under a longitudinal study design. This model detects latent topics that describe examinee responses and characterizes the change in the use of topics over time. The proposed model is illustrated by a real data analysis from an instructional intervention study on teaching science inquiry skills to middle school students. A simulation study is also provided to evaluate the growth curve topic model under several conditions.

Introduction

Many assessments from large-scale standardized assessments to classroom assessments include constructed response (CR) items in addition to selected response items to measure students' knowledge that multiple choice items are less suited to measure (Attali, 2014; Bennett, 1991; Rodriguez, 2003; Wang, Song, Wang, & Wolfe, 2017). Student responses in these types of items have been found to have useful information regarding student learning and achievement (e.g., Buxton, Allexsaht-Snider, Aghasaleh, Kayumova, Kim, Choi, & Cohen, 2014; Crossley, Salsbury, & McNamara, 2015; Furtak & Ruiz-Primo, 2008). The focus of this paper is particularly on text data that are collected longitudinally over multiple time points.

In social and behavioral sciences research, including educational research, the measurement of change has received considerable attention (Bollen & Curran, 2006; Rogosa, Brandt, & Zimowski, 1982). Research on measurement of change, however, has focused almost entirely on development of models and analysis of data that consist of numeric responses (e.g., McArdle, 2009; Muthén & Curran, 1997; Singer, 1998). When the original data are text, the textual responses are typically first graded using a scoring rubric, and then the scores are analyzed (e.g., Crossley, Salsbury, & McNamara, 2009; Daller, Turlik, & Weir, 2013).

In this study, we propose a new approach that incorporates a method of text analysis called topic modeling and growth curve modeling to model change in use of language as a response to instruction. We revisit the original scored textual data from an instructional intervention with an eye to detecting additional information about the ways in which examinees change their use of language over the course of an intervention. The new model provides an exploratory method that can summarize written responses collected over multiple time points.

Topic modeling is a statistical approach that has been widely used for analyzing large amounts of text data. Topic models assume that a document is composed of a mixture of topics (Blei, 2012). Topic models treat a document as a set of words but usually ignore the order of words. This is referred to as the bag-of-words assumption. This topic modeling framework stems from the probabilistic latent semantic indexing method, also referred to as the aspect model (Hofmann, 1999; Hoffmann, 2001). The aspect model did not specify from where the topic proportions of documents derive. This makes it difficult to make inferences for a new document. Blei, Ng, and Jordan (2003) extended this model by introducing a Dirichlet prior on the topic proportion variable, and developed latent Dirichlet allocation (LDA). The LDA model is designed to detect the latent thematic structure in a corpus (i.e., a set of documents). Each topic

is represented as a fixed vocabulary associated with probabilities of the appearance of corresponding words in a document given the topic.

Topic modeling has been widely used across many disciplines including education (e.g., Chen, Yu, Zhang, & Yu, 2016; Johri, Wang, Liu, & Madhavan , 2011; Kim, Kwak, Cardozo-Gaibisso, Buxton, & Cohen, 2017; Reich, Tingley, Leder-Luis, Roberts, & Stewart, 2015), psychology (e.g., Socher, Gershman, Sederberg, Norman, Perotte, & Blei, 2009), political science (e.g., Lauderdale & Clark, 2014; Lucas, Nielsen, Roberts, Stewart, Storer, & Tingley, 2015), health science (e.g., Paul & Dredze, 2011), and bioinformatics (e.g., Liu, Tang, Dong, Yao, & Zhou, 2016). In education, topic modeling has been used to analyze student-generated texts. Chen et al. (2016) analyzed pre-service teachers' journals using LDA to detect the latent thematic structure discussed in students' journals. Kim et al. (2017) analyzed middle school students' responses to CR items using LDA to detect topics in students' answers and how the topic proportions or topic usage changed from pretest to posttest. Johri et al. (2011) analyzed journal and conference papers in engineering education using LDA to investigate how topics changed over a period of nine years. Reich et al. (2015) analyzed student writings generated in a massive open online course setting using the structural topic model (Roberts, Stewart, Tingley, Lucas, Leder-Luis, Gadarian, Albertson, & Rand, 2014; Roberts, Stewart, & Airoldi, 2016).

In this study, we focus on change over multiple time points in use of the latent topics in the text of students' responses to CR items. Several topic models incorporate the time at which documents were created or published. Blei and Lafferty (2006) proposed a dynamic topic model that allows topics to evolve over time. Wang, Blei, and Heckerman (2012) introduced a dynamic topic model for continuous time. Gerrish and Blei (2010) introduced the document influence model by using a dynamic topic model. This document influence model allows important

documents to influence future topics more significantly than the other documents in the corpus. Glynn, Tokdar, Howard, and Banks (2019) extended the dynamic topic model and proposed a dynamic linear topic model. These models, however, are not necessarily developed for situations in which the same individuals produce data over multiple times. In this study, a new approach is proposed that combines growth curve modeling and topic modeling to analyze text data collected under a longitudinal study design.

Growth curve modeling (Bollen & Curran, 2006) is a statistical method for analyzing data that have repeated measurements. Growth curve models estimate individuals' change over time and inter-individual variability in that change (Bollen & Curran, 2006; Curran & Muthén, 1999; Curran, Obeidat, & Losardo, 2010). These models have been widely used in the social and behavioral sciences to characterize and understand change in subjects (e.g., Cillessen & Mayeux, 2004; Curran, Harford, & Muthén, 1996; McCoach, O'Connell, Reis, & Levitt, 2006; Owens & Shaw, 2003; Williams, Conger, & Blozis, 2007). In the method proposed here, the topic modeling component captures topics in the texts of students' responses to CR items, and the growth curve modeling component characterizes changes in topic proportions over time.

This study offers three main contributions. First, we introduce a new approach to analyzing text data collected over multiple time points. Longitudinal settings are often used in educational studies to capture students' behavioral changes, such as response to instruction. This study provides a statistical method for analyzing text data that were collected over multiple time points and provides information that differs from what we can obtain from the numeric scores given to students' responses to CR items. Second, we provide an estimation method to estimate the model parameters. This method uses a Gibbs sampling algorithm (Geman & Geman, 1984) and a Metropolis-Hastings algorithm (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth,

Teller, & Teller, 1953) for posterior inferences. The estimation method is described in detail in the Methods section. Third, we provide an empirical data analysis to illustrate the use of the proposed model. The data were from an NSF-funded study that focused on guiding teachers in how to teach their students use of academic language for understanding science inquiry practices.

The remainder of this study is organized as follows. First, we describe the proposed model and the estimation process. Next, we provide an analysis of empirical data using a dataset collected across four time points from middle school students. We used the proposed model to investigate the latent topics underlying students' written responses and then to characterize the change in use of topics over time. A simulation study follows which evaluates the model's performance under practical testing conditions.

Methods

This study incorporates latent Dirichlet allocation (LDA; Blei et al., 2003) and a latent growth curve model (Bollen & Curran, 2006). LDA is designed to extract latent topics from textual data. Latent growth curve models are designed for describing changes in response over time. The changes can be described in various ways, such as linear or quadratic curves. The proposed growth curve topic model is developed by combining LDA and a linear growth curve model in order to simultaneously detect the latent topic structure and model changes in use of topic proportions in a set of longitudinal text data. As described in this study, one possible application of this model is for situations in which examinees take a constructed response test several times, and the change in response over time is the main interest.

Model Description

The growth curve topic model consists of two components: 1) the topic modeling component, which identifies topics and topic proportions from documents, and 2) the structural

modeling component, which describes systemic changes in topic proportions. In other words, in the growth curve topic model, documents across T time points are modeled using the topic modeling component, and the changes in topic proportions across T time points are described using the structural modeling component.

Topic model. To describe the topic modeling component, the following terminology and notation are used.

1) Words and documents: words are observed data. Each document i at time point t consists of n_{it} number of words which is represented as $\boldsymbol{w}_{i,t} = \left(w_{1,i,t}, w_{2,i,t}, \dots, w_{n_{it},i,t}\right)^T$ for $i = 1, \dots, D$ and $t = 1, \dots, T$.
2) Topics: Topic modeling assumes that texts have latent thematic structures, which are referred to as topics. Topics are viewed as multinomial distributions over the vocabulary of the corpus with associated probabilities $\boldsymbol{\gamma}_k = (\gamma_{k1}, \dots, \gamma_{kV})^T$ for topic k, where V is the number of unique words in the corpus. Words associated with high probabilities in a multinomial distribution characterize the corresponding topic.
3) Topic assignments: Each word $w_{n,i,t}$ is assumed to be from one of the topics. The topic assignment for $w_{n,i,t}$ is indicated by $z_{n,i,t}$.
4) Topic proportions: Each document i at time t has a vector of topic proportions $\boldsymbol{\theta}_{i,\cdot,t} = \left(\theta_{i,1,t}, \dots, \theta_{i,K,t}\right)^T$, where K is the number of topics, and $\sum_{k=1}^{K} \theta_{i,k,t} = 1$. A topic is assigned to each word in document i at time t using a multinomial distribution with $\boldsymbol{\theta}_{i,\cdot,t}$. In this study, the natural parameterization of a multinomial distribution is used:

$$\eta_{i,k,t} = \log \frac{\theta_{i,k,t}}{\theta_{i,K,t}}.$$

The last topic (topic K) is taken as the reference topic. According to this equation, $\eta_{i,K,t}$ is 0 for all $i = 1, \dots, D$ and $t = 1, \dots, T$.

Suppose that a corpus consists of V unique words, which are indexed by v. The generative process assumed by the topic modeling component is as follows:

1) Choose $\boldsymbol{\gamma}_k \sim Dirichlet(\boldsymbol{\beta})$ for $k = 1, \dots, K$, where $\boldsymbol{\beta}$ is a $V \times 1$ vector of Dirichlet parameters, and $\boldsymbol{\gamma}_k = (\gamma_{k1}, \dots, \gamma_{kV})^T$ is a $V \times 1$ vector with $\sum_{v=1}^{V} \gamma_{kv} = 1$. Each element of $\boldsymbol{\gamma}_k$ indicates the probability of a corresponding word appearing in a document under topic k.

For each document i:

2) Choose $\boldsymbol{\eta}_{i,k,\cdot} = \left(\eta_{i,k,1}, \dots, \eta_{i,k,T}\right)^T \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Phi}_{\mathrm{k}})$ for $k = 1, \dots, K-1$.
3) For each word $(w_{n,i,t})$ in document i at time t:
    a. Choose a topic $z_{n,i,t} \sim Multinomial\left(\pi\left(\boldsymbol{\eta}_{i,\cdot,t}\right)\right)$, where $\boldsymbol{\eta}_{i,\cdot,t} = \left(\eta_{i,1,t}, \dots, \eta_{i,K,t}\right)^T$.
    b. Choose a word $w_{n,i,t} \sim Multinomial\left(\boldsymbol{\gamma}_{z_{n,i,t}=k}\right)$.

The π maps the natural parameterization of topic proportions to the mean parameterization,

$$\pi\left(\boldsymbol{\eta}_{i,\cdot,t}\right) = \left(\frac{\exp\left(\eta_{i,1,t}\right)}{\sum_{k=1}^{K}\exp\left(\eta_{i,k,t}\right)}, \dots \frac{\exp\left(\eta_{i,K,t}\right)}{\sum_{k=1}^{K}\exp\left(\eta_{i,k,t}\right)}\right)^T = \left(\theta_{i,1,t}, \dots, \theta_{i,K,t}\right)^T.$$

The detailed specification of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Phi}_k$ is described below, in the structural model section.

Structural model. Given time t and topic k, individual i's topic k proportion is modeled using a linear growth curve model as follows:

$$\eta_{i,k,t} = \xi_{a,i,k} + (t-1)\xi_{b,i,k}, \qquad (4.1)$$

where $\xi_{a,i,k}$ is an intercept for individual i, and $\xi_{b,i,k}$ is a slope for individual i. Equation (4.1) can be specified in a matrix form using $\boldsymbol{\eta}_{i,k,\cdot} = \left(\eta_{i,k,1}, \ldots, \eta_{i,k,T}\right)^T$:

$$\boldsymbol{\eta}_{i,k,\cdot} = \boldsymbol{\Lambda}\boldsymbol{\xi}_{i,k}, \qquad (4.2)$$

where

$$\boldsymbol{\Lambda}^{\mathrm{T}} = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 0 & 1 & \ldots & T-1 \end{pmatrix}$$

and $\boldsymbol{\xi}_{i,k} = \left(\xi_{a,i,k}, \xi_{b,i,k}\right)^T$. We assume $\boldsymbol{\xi}_{i,k}$ to follow a normal distribution with a mean vector of $\boldsymbol{\mu}_{\xi_k} = \left(\mu_{\xi_{ak}}, \mu_{\xi_{bk}}\right)^T$ and a variance-covariance matrix of $\boldsymbol{\Psi}_k$. Then, $\boldsymbol{\eta}_{i,k,\cdot}$ follows $N(\boldsymbol{\mu}_k, \boldsymbol{\Phi}_k)$, with $\boldsymbol{\mu}_k = \boldsymbol{\Lambda}\boldsymbol{\mu}_{\xi_k}$ and $\boldsymbol{\Phi}_k = \boldsymbol{\Lambda}\boldsymbol{\Psi}_k\boldsymbol{\Lambda}^T$.

The mean of $\xi_{a,i,k}$ indicates a mean value for the log odds of using topic k rather than topic K at time 1, and the mean of $\xi_{b,i,k}$ indicates the rate of change in the log odds, when the time point increases by one unit. A positive mean of $\xi_{b,i,k}$ indicates that the log odds of using topic k rather than topic K tends to increase in documents over time, and a negative mean of $\xi_{b,i,k}$ indicates that the log odds tends to decrease in documents over time.

Estimation

***Markov chain Monte Carlo algorithm*.** Markov chain Monte Carlo (MCMC) methods are useful for simulating complex multivariate distributions. An MCMC method generates random samples from a multivariate distribution of interest, and makes inferences about the distribution using the features of the random samples. Gibbs sampling (Gelfand & Smith, 1990) is one of the MCMC methods for generating samples from the posterior distribution of parameters of interest. In the Gibbs sampling, the parameter vector is divided into a number of subvectors. These subvectors are sampled one at a time, conditioning on the value of all the other parameters (Gelman, Carlin, Stern, Dunson, Vehtari, & Rubin, 2013). When the full conditional

distribution for some parameters is not in a known form, we can obtain samples from the full conditional distribution using a Metropolis-Hastings step within Gibbs sampling. The Metropolis-Hastings algorithm (Chib & Greenberg, 1995) is also one of the MCMC methods, and it is useful when direct sampling from a target distribution is difficult. This algorithm uses a proposal distribution from which it is easy to sample. The proposal distribution generates candidate samples, and these are either accepted or rejected for samples from the approximation of the target distribution.

In this study, we used Gibbs sampling with a Metropolis-Hastings step for the posterior inference of model parameters. Data augmentation approach (Tanner & Wong, 1987) was also used. Data augmentation is an MCMC technique for constructing an iterative algorithm by adding unobserved data (Van Dyk & Meng, 2001). We augment the observed document data with topic assignments ($z_{n,i,t}$), and the vector of random intercept and slope ($\boldsymbol{\xi}_{i,k}$).

Joint distribution. Suppose $\boldsymbol{\xi}_{\cdot,k} = \{\boldsymbol{\xi}_{1,k}, \boldsymbol{\xi}_{2,k} \dots, \boldsymbol{\xi}_{D,k}\}$, $\boldsymbol{\xi} = \{\boldsymbol{\xi}_{\cdot,1}, \boldsymbol{\xi}_{\cdot,2}, \dots, \boldsymbol{\xi}_{\cdot,K-1}\}$, $\boldsymbol{\mu}_{\xi} = \{\boldsymbol{\mu}_{\xi_1}, \boldsymbol{\mu}_{\xi_2}, \dots \boldsymbol{\mu}_{\xi_{K-1}}\}$, $\boldsymbol{\Psi} = \{\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2, \dots, \boldsymbol{\Psi}_{K-1}\}$, $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \dots, \boldsymbol{\gamma}_K\}$, $\boldsymbol{w}_i = \{\boldsymbol{w}_{i,1}, \boldsymbol{w}_{i,2}, \dots, \boldsymbol{w}_{i,T}\}$, $\boldsymbol{w} = \{\boldsymbol{w}_1, \dots \boldsymbol{w}_D\}$, $\boldsymbol{z}_{i,t} = \{z_{1,i,t}, z_{2,i,t}, \dots, z_{n_{it},i,t}\}$, $\boldsymbol{z}_i = \{\boldsymbol{z}_{i,1}, \boldsymbol{z}_{i,2}, \dots, \boldsymbol{z}_{i,T}\}$ and $\boldsymbol{z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2 \dots, \boldsymbol{z}_D\}$. The joint distribution of $\boldsymbol{\xi}$, $\boldsymbol{\mu}_{\xi}$, $\boldsymbol{\Psi}$, $\boldsymbol{\gamma}$, $\boldsymbol{z}$, and $\boldsymbol{w}$ is

$$p(\boldsymbol{\xi}, \boldsymbol{\mu}_{\xi}, \boldsymbol{\Psi}, \boldsymbol{\gamma}, \boldsymbol{w}, \boldsymbol{z}) = p(\boldsymbol{\gamma})p(\boldsymbol{w}|\,\boldsymbol{\gamma}, \boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{\xi}) \prod\nolimits_{k=1}^{K-1} p(\boldsymbol{\xi}_{\cdot,k}|\boldsymbol{\mu}_{\xi_k}, \boldsymbol{\Psi}_k)p(\boldsymbol{\mu}_{\xi_k})p(\boldsymbol{\Psi}_k)$$

$$= \left[\prod_{k=1}^{K} p(\boldsymbol{\gamma}_k)\right]\left[\prod_{i=1}^{D}\prod_{t=1}^{T}\{p(\boldsymbol{w}_{i,t}|\boldsymbol{z}_{i,t}, \boldsymbol{\gamma})p(\boldsymbol{z}_{i,t}|\boldsymbol{\xi})\}\right]\left[\prod_{k=1}^{K-1}\left\{\left(\prod_{i=1}^{D} p(\boldsymbol{\xi}_{i,k}|\boldsymbol{\mu}_{\xi_k}, \boldsymbol{\Psi}_k)\right)p(\boldsymbol{\mu}_{\xi_k})p(\boldsymbol{\Psi}_k)\right\}\right]. \quad (4.3)$$

Full conditional distributions. The full conditional distributions of $\boldsymbol{\gamma}_k$, $\boldsymbol{\xi}_{i,k}$, $\boldsymbol{\mu}_{\xi_k}$, $\boldsymbol{\Psi}_k$, and $z_{n,i,t}$ are specified below.

At the $(m+1)$th iteration,

1) Update $\boldsymbol{\gamma}_k$ $(k = 1, \dots, K)$

$\boldsymbol{\gamma}_k$ is a $V \times 1$ vector of word probabilities for topic k. Here, $\boldsymbol{\gamma}_k \sim Dirichlet(\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a $V \times 1$ vector of Dirichlet distribution parameters, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_V)^T$. From Equation (4.3), the full conditional distribution of $\boldsymbol{\gamma}_k$ is

$$p(\boldsymbol{\gamma}_k|\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\beta})$$

$$\propto p(\boldsymbol{\gamma}_k|\boldsymbol{\beta})p(\boldsymbol{w}|\boldsymbol{z} = k, \boldsymbol{\gamma}_k)$$

$$\propto \prod_{v=1}^{V} \gamma_{kv}^{\beta_v - 1} \prod_{v=1}^{V} \gamma_{kv}^{N_{kv}}$$

$$= \prod_{v=1}^{V} \gamma_{kv}^{N_{kv} + \beta_v - 1},$$

where N_{kv} is the number of occurrences that word v is assigned to topic k across all documents. Thus, we can use the following Dirichlet distribution to obtain the $(m + 1)$th sample of $\boldsymbol{\gamma}_k$:

$$\boldsymbol{\gamma}_k^{(m+1)} \sim Dirichlet(\boldsymbol{N}_k + \boldsymbol{\beta}),$$

where $\boldsymbol{N}_k = (N_{k1}, \dots, N_{kV})^T$.

2) Update $\boldsymbol{\xi}_{i,k}$ $(i = 1, \dots, D; k = 1, \dots, K - 1)$

From Equation (4.3), the full conditional distribution of $\boldsymbol{\xi}_{i,k}$ is

$$p\left(\boldsymbol{\xi}_{i,k}|\, \boldsymbol{z}, \boldsymbol{\mu}_{\xi_k}, \boldsymbol{\Psi}_k, \boldsymbol{\xi}_{-(i,k)}\right)$$

$$\propto p\left(\boldsymbol{\xi}_{i,k}|\boldsymbol{\mu}_{\xi_k}, \boldsymbol{\Psi}_k\right) \prod_{t=1}^{T} p\left(\boldsymbol{z}_{i,t}|\boldsymbol{\xi}\right)$$

$$\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\xi}_{i,k} - \boldsymbol{\mu}_{\xi_k}\right)^T \boldsymbol{\Psi}_k^{-1}\left(\boldsymbol{\xi}_{i,k} - \boldsymbol{\mu}_{\xi_k}\right)\right\} \prod_{t=1}^{T} \prod_{k=1}^{K} \theta_{i,k,t}^{x_{i,k,t}}, \qquad (4.4)$$

where $\boldsymbol{\xi}_{-(i,k)}$ is equal to $\boldsymbol{\xi}$ without $\boldsymbol{\xi}_{i,k}$ element, $x_{i,k,t}$ is the total number of words that were assigned to topic k for document i at time t, and

$$\theta_{i,k,t} = \pi\left(\boldsymbol{\lambda}_t \boldsymbol{\xi}_{i,k}\right) = \frac{\exp\left(\boldsymbol{\lambda}_t \boldsymbol{\xi}_{i,k}\right)}{\sum_{k=1}^{K} \exp\left(\boldsymbol{\lambda}_t \boldsymbol{\xi}_{i,k}\right)},$$

in which $\boldsymbol{\lambda}_t$ is the tth row of $\boldsymbol{\Lambda}$. Since direct sampling of $\boldsymbol{\xi}_{i,k}$ from Equation (4.4) is difficult, we additionally employ a Metropolis-Hastings step to update $\boldsymbol{\xi}_{i,k}$, which is implemented as follows:

a) Sample $\boldsymbol{\xi}_{i,k}^{(new)}$ from a proposal distribution, $N\left(\boldsymbol{\xi}_{i,k}^{(m)}, c\boldsymbol{I}\right)$, where c is a tuning parameter, and $\boldsymbol{I}$ is an identity matrix.

b) Calculate the ratio of densities,

$$r = \frac{\exp\left\{-\frac{1}{2}\left(\boldsymbol{\xi}_{i,k}^{(new)} - \boldsymbol{\mu}_{\xi_k}\right)^T \boldsymbol{\Psi}_k^{-1}\left(\boldsymbol{\xi}_{i,k}^{(new)} - \boldsymbol{\mu}_{\xi_k}\right)\right\} \prod_{t=1}^{T}\prod_{k=1}^{K} \pi\left(\boldsymbol{\lambda}_t \boldsymbol{\xi}_{i,k}^{(new)}\right)^{x_{i,k,t}}}{\exp\left\{-\frac{1}{2}\left(\boldsymbol{\xi}_{i,k}^{(m)} - \boldsymbol{\mu}_{\xi_k}\right)^T \boldsymbol{\Psi}_k^{-1}\left(\boldsymbol{\xi}_{i,k}^{(m)} - \boldsymbol{\mu}_{\xi_k}\right)\right\} \prod_{t=1}^{T}\prod_{k=1}^{K} \pi\left(\boldsymbol{\lambda}_t \boldsymbol{\xi}_{i,k}^{(m)}\right)^{x_{i,k,t}}}.$$

c) Assign $\boldsymbol{\xi}_{i,k}^{(m+1)} = \boldsymbol{\xi}_{i,k}^{(new)}$ with probability of $min(r, 1)$, otherwise assign $\boldsymbol{\xi}_{i,k}^{(m+1)} = \boldsymbol{\xi}_{i,k}^{(m)}$.

3) Update $\boldsymbol{\mu}_{\xi_k}$ $(k = 1, \ldots, K-1)$

Suppose that the prior distribution of $\boldsymbol{\mu}_{\xi_k}$ is $\boldsymbol{\mu}_{\xi_k} \sim N(\boldsymbol{\mu}_{0k}, \boldsymbol{\Sigma}_{0k})$. From Equation (4.3), the full conditional distribution of $\boldsymbol{\mu}_{\xi_k}$ is

$$p\left(\boldsymbol{\mu}_{\xi_k} | \boldsymbol{\xi}_{\cdot,k}, \boldsymbol{\Psi}_k\right)$$

$$\propto p\left(\boldsymbol{\mu}_{\xi_k}\right) p\left(\boldsymbol{\xi}_{\cdot,k} | \boldsymbol{\mu}_{\xi_k}, \boldsymbol{\Psi}_k\right)$$

$$\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\mu}_{\xi_k} - \boldsymbol{\mu}_{0k}\right)^T \boldsymbol{\Sigma}_{0k}^{-1}\left(\boldsymbol{\mu}_{\xi_k} - \boldsymbol{\mu}_{0k}\right)\right\} \prod_{i=1}^{D} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\xi}_{i,k} - \boldsymbol{\mu}_{\xi_k}\right)^T \boldsymbol{\Psi}_k^{-1}\left(\boldsymbol{\xi}_{i,k} - \boldsymbol{\mu}_{\xi_k}\right)\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\mu}_{\xi_k}-\boldsymbol{\mu}_{0k}\right)^T\boldsymbol{\Sigma}_{0k}^{-1}\left(\boldsymbol{\mu}_{\xi_k}-\boldsymbol{\mu}_{0k}\right)-\frac{1}{2}\sum_{i=1}^{D}\left(\boldsymbol{\xi}_{i,k}-\boldsymbol{\mu}_{\xi_k}\right)^T\boldsymbol{\Psi}_k^{-1}\left(\boldsymbol{\xi}_{i,k}-\boldsymbol{\mu}_{\xi_k}\right)\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\mu}_{\xi_k}^T\boldsymbol{\Sigma}_{0k}^{-1}\boldsymbol{\mu}_{\xi_k}-\boldsymbol{\mu}_{\xi_k}^T\boldsymbol{\Sigma}_{0k}^{-1}\boldsymbol{\mu}_{0k}-\boldsymbol{\mu}_{0k}^T\boldsymbol{\Sigma}_{0k}^{-1}\boldsymbol{\mu}_{\xi_k}\right.\right.$$

$$\left.\left.-\sum_{i=1}^{D}\left(\boldsymbol{\xi}_{i,k}^T\boldsymbol{\Psi}_k^{-1}\boldsymbol{\mu}_{\xi_k}+\boldsymbol{\mu}_{\xi_k}^T\boldsymbol{\Psi}_k^{-1}\boldsymbol{\xi}_{i,k}-\boldsymbol{\mu}_{\xi_k}^T\boldsymbol{\Psi}_k^{-1}\boldsymbol{\mu}_{\xi_k}\right)\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\mu}_{\xi_k}^T(D\boldsymbol{\Psi}_k^{-1}+\boldsymbol{\Sigma}_{0k}^{-1})\boldsymbol{\mu}_{\xi_k}-\boldsymbol{\mu}_{\xi_k}^T\left(\sum_{i=1}^{D}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\xi}_{i,k}+\boldsymbol{\Sigma}_{0k}^{-1}\boldsymbol{\mu}_{0k}\right)\right.\right.$$

$$\left.\left.-\left(\sum_{i=1}^{D}\boldsymbol{\xi}_{i,k}^T\boldsymbol{\Psi}_k^{-1}+\boldsymbol{\mu}_{0k}^T\boldsymbol{\Sigma}_{0k}^{-1}\right)\boldsymbol{\mu}_{\xi_k}\right]\right\}.$$

Thus, we can use the following normal distribution to obtain the $(m+1)$th sample of $\boldsymbol{\mu}_{\xi_k}$:

$$\boldsymbol{\mu}_{\xi_k}^{(m+1)} \sim N\left(m\left(\boldsymbol{\mu}_{\xi_k}\right), V\left(\boldsymbol{\mu}_{\xi_k}\right)\right),$$

where

$$m\left(\boldsymbol{\mu}_{\xi_k}\right) = V\left(\boldsymbol{\mu}_{\xi_k}\right)\left(\sum_{i=1}^{D}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\xi}_{i,k}+\boldsymbol{\Sigma}_{0k}^{-1}\boldsymbol{\mu}_{0k}\right),$$

and

$$V\left(\boldsymbol{\mu}_{\xi_k}\right) = (D\boldsymbol{\Psi}_k^{-1}+\boldsymbol{\Sigma}_{0k}^{-1})^{-1}.$$

4) Update $\boldsymbol{\Psi}_k$ $(k = 1, \ldots, K-1)$

Suppose that $\boldsymbol{\Psi}_k$ follows an inverse-Wishart distribution with degrees of freedom v_0 and scale matrix $\boldsymbol{S}_0$, $\boldsymbol{\Psi}_k \sim IW(v_0, \boldsymbol{S}_0)$. Then, the full conditional distribution of $\boldsymbol{\Psi}_k$ also

follows an inverse-Wishart distribution. Thus, we can obtain the $(m+1)$th sample of $\boldsymbol{\Psi}_k$ from

$$\boldsymbol{\Psi}_k^{(m+1)} \sim IW\left(v_p, \boldsymbol{S}_p\right),$$

where

$$v_p = D + v_0$$

and

$$\boldsymbol{S}_p = \boldsymbol{S}_0 + \sum_{i=1}^{D} \left(\boldsymbol{\xi}_{i,k} - \boldsymbol{\mu}_{\xi_k}\right)\left(\boldsymbol{\xi}_{i,k} - \boldsymbol{\mu}_{\xi_k}\right)^T.$$

5) Update $z_{n,i,t}$ $(n = 1, \ldots, n_{it})$ for each individual i and time t

The full conditional distribution of $z_{n,i,t}$ is

$$p\left(z_{n,i,t} = k | w_{n,i,t} = v, \boldsymbol{\gamma}_k, \boldsymbol{\xi}\right)$$

$$\propto p\left(z_{n,i,t} = k | \boldsymbol{\xi}\right) p\left(w_{n,i,t} = v | z_{n,i,t} = k, \boldsymbol{\gamma}_k\right)$$

$$\propto \frac{\exp\left(\boldsymbol{\lambda}_t \boldsymbol{\xi}_{i,k}\right)}{\sum_{k=1}^{K} \exp\left(\boldsymbol{\lambda}_t \boldsymbol{\xi}_{i,k}\right)} \gamma_{kv}$$

$$\propto e^{\boldsymbol{\lambda}_t \xi_{i,k}} \gamma_{kv}.$$

Thus, the $(m+1)$th value of $z_{n,i,t}$ can be obtained from a multinomial distribution with a vector of unnormalized probabilities $e^{\boldsymbol{\lambda}_t \xi_{i,k}} \gamma_{kv}$ $(k = 1, \ldots, K)$.

Empirical Data Analysis

In this section, an empirical data analysis using students' responses to constructed response (CR) items is described to illustrate the use of the growth curve topic model. We first describe the data used for this study and then the LDA analysis results used to determine the number of topics in the data. Next, the estimation process of the growth curve topic model for

this data is described, and results of the growth curve topic model analysis are reported. Finally, the model fit is evaluated using Bayesian posterior predictive checks.

Data

The data for this study consisted of CR responses from a sample of 243 middle grades students that participated in an instructional intervention in an NSF-funded host study. The study focused on how to teach middle grades students' science inquiry practices. As a part of this project, a science assessment was designed to measure students' understanding of science inquiry practices and use of academic language. Two parallel forms, Form A and Form B, of the assessment were administered to middle school students on two occasions in each of the two contiguous academic years of the study. The assessment was administered as a pre-test at the beginning of each school year and as a post-test at the end of each school year. For the pre-test, one of the two forms was randomly given to each student. For the post-test, each student received the opposite form from the one they took for the pre-test. The instructional intervention lasted one academic semester, and focused on teaching science inquiry practices. The intervention was taught to teachers following the pre-test. Teachers than taught it to their students. The post-test was administered after the completion of the instructional intervention.

The two forms of the assessment consisted of six science experiment scenarios. Each scenario asked students to respond to from two to four constructed response items and was designed to measure the use of independent and dependent variables, construction of hypotheses, and cause and effect. Students' responses were written in separate answer documents distributed along with the test booklets. In this study, we used students' responses from Form A to track the use of topics over time.

In this empirical data analysis, each student's responses in their respective answer document to all six scenarios were considered as a single document. Topic model analysis requires the response data to be pre-processed before conducting the analysis. In pre-processing the data, words were converted into their basic roots, and stopwords, such as *a, and, the* were removed. In addition, words that appeared less than 5 times in documents were excluded. Also, students who had fewer than 10 words in their answer documents were excluded. As a result, the data for this study consisted of 493 documents with 559 unique words from 243 students. (The total number of documents was smaller than $243 \times 4 = 972$ because we only used documents from the Form A assessment.) The average length of documents was 51.77 words. Table 4.1 shows detailed descriptive information regarding the student response data.

Investigating the Number of Topics

In topic modeling research, there is no single agreed upon best model selection criterion for determining the number of topics. Blei et al. (2003) used perplexity to determine the number of topics. Griffiths and Steyvers (2004) used the log likelihood of the data in order to find the appropriate number of topics. Chen et al. (2016) used a stability score to find the number of topics that had the most stability. Lauderdale and Clark (2014) used the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) to determine the number of topics. However, Chang, Gerrish, Wang, Boyed-Graber, and Blei (2009) showed that statistical model indexes for choosing the number of topics in a corpus and human judgement did not agree. Grimmer and Stewart (2013) suggested including use of human judgement in addition to statistical indexes when selecting the number of topics for a corpus.

In this study, we considered a Bayesian measures of model complexity, DIC and AICM (Raftery, Newton, Satagopan, & Krivitsky, 2007) to determine the number of topics in the

student responses. These indexes can be useful when a posterior distribution of parameters is obtained from a MCMC algorithm in that DIC and AICM can be readily obtained from the MCMC output. We also examined topics by comparing topics that were obtained from different topic models.

The number of topics was determined by using LDA. This is equivalent to determining the number of topics by using the topic modeling component of the growth curve topic model. The detailed equations for DIC and AICM for LDA are given in Appendix C. An exploratory LDA analysis was fit by estimating models with two topics, three topics, etc., up to an LDA model with six topics to the student response data. The maximum possible number of topics was set to be six as the student responses were restricted to the six science scenarios. Previous research on similar data (Kim et al., 2017) suggested that this limited the possible number of topics that could appear in the responses.

Table 4.2 presents DIC and AICM values for the five LDA models. A smaller DIC value and a larger AICM value indicate a better model fit. The DIC values indicate that the LDA model with six topics was the best fitting model, and the AICM values indicate that the LDA model with three topics was the best fitting model. After examining the topics from the five different LDA models carefully, the three-topic model was determined to be the most appropriate for summarizing the student response data. This was because, when more than three topics were assumed, some topics were similar to another.

Table 4.3 presents the 15 words with the highest posterior mean estimates of γ_{kv} for each of the three topics. For each word, γ_{kv} indicates the posterior mean estimate of the probability that the word appears in a document when the word had assignment to the corresponding topic. Words in the Topic 1 column are mainly related to everyday language, words in the Topic 2

column are generally related to discipline specific language, and words in Topic 3 column are generally concerned with general academic language.

Table 4.4 presents the average of posterior means of topic proportions for documents from four time points. The proportions in this table indicate that topic proportions for Topic 1 and Topic 3 decreased from Time 1 to Time 4 in general, and topic proportions for Topic 2 tended to increase from Time 1 to Time 4. In the next section, we characterize this change in topic proportions using the growth curve topic model.

Growth Curve Topic Model Analysis

Estimation. The growth curve topic model with $K = 3$ was fit to the student response data. In this analysis, Topic 3 was assigned as the reference topic. As described in the data section, students' responses from one of the forms, Form A, were analyzed in this study. Because of the study design, each student had missing responses (documents) for at most two of the four time points. This type of missing responses can be considered as planned missingness, and the missing at random (MAR) assumption can be hold (Bollen & Curran, 2006; Enders, 2010; Rhemtulla & Hancock, 2016; Schafer & Graham, 2002). Missing responses lead to missing topic proportions in the structural model component. In order to handle the missing responses, words in missing documents and missing topic assignments were treated as latent variables and were sampled during the MCMC sampling process described below. The lengths of missing documents for each time point were randomly sampled from a Poisson distribution in a way that the mean of the sampled lengths was similar to the average document length obtained from observed documents. The Gibbs sampling algorithm including the missing data treatment used for analyzing the student responses is as follows.

1) Update $\boldsymbol{\gamma}_k$

   A prior distribution on $\boldsymbol{\gamma}_k$ was placed as $\boldsymbol{\gamma}_k \sim Dirichlet(\boldsymbol{\beta})$ with $\boldsymbol{\beta} = \mathbf{1}_V$, where $\mathbf{1}_V$ is a $V \times 1$ vector consisting of all 1s.

2) Update $\boldsymbol{\xi}_{i,k}$

   The tuning parameter c for each topic was chosen in a way that the average acceptance rate was approximately 0.4.

3) Update $\boldsymbol{\mu}_{\xi_k}$

   A prior distribution on $\boldsymbol{\mu}_{\xi_k}$ was placed as $\boldsymbol{\mu}_{\xi_k} \sim N(\boldsymbol{\mu}_{0k}, \boldsymbol{\Sigma}_{0k})$ with $\boldsymbol{\mu}_{0k} = (0,0)^T$ and $\boldsymbol{\Sigma}_{0k} = 100\boldsymbol{I}_2$, where $\boldsymbol{I}_2$ is a 2×2 identity matrix.

4) Update $\boldsymbol{\Psi}_k$

   A prior distribution was placed on $\boldsymbol{\Psi}_k$ as $\boldsymbol{\Psi}_k \sim IW(v_0, \boldsymbol{S}_0)$ with $v_0 = 3$ and $\boldsymbol{S}_0 = \boldsymbol{I}_2$.

5) Update $z_{n,i,t}$

6) For each of the missing documents:

   a. Sample topic assignments from the multinomial distribution described in the algorithm section using sampled missing words.

   b. Sample missing words using sampled topic assignments and updated $\boldsymbol{\gamma}_k$.

The MCMC chain was run for a total of 50,000 iterations. Convergence was determined using Geweke's (Geweke, 1992) convergence diagnostic as implemented in the R package *coda* (Plummer, Best, Cowles, & Vines, 2006). The Geweke's diagnostic values for all parameters were within the range of -2 and 2 after 15,000 iterations, which indicated that the chains had converged. These first 15,000 iterations were discarded as burn-in. The subsequent chain was thinned to every 5th iterations because of limitations in computer memory and storage. Figures 4.1 and 4.2 presents the trace plots of $\boldsymbol{\mu}_{\xi_k}$and $\boldsymbol{\Psi}_k$ for the thinned chain. The first row of the

figures are the trace plots for $k = 1$, and the second row of the figures are the trace plots for $k = 2$. The trace plots for $\boldsymbol{\mu}_{\xi_k}$ and $\boldsymbol{\Psi}_k$ appeared to be stabilized after the burn-in period.

Results. Table 4.5 presents the 15 words with the highest posterior mean estimates of γ_{kv}. It can be seen that some words such as *change* and *cause* appeared among the 15 words of more than one topic in this table. This is because words can appear in all topics. The 15 words in the Topic 1 column mainly relate to everyday language. The 15 words in the Topic 2 column mainly relate to discipline-specific language. The 15 words in the Topic 3 column mostly relate to general technical vocabulary. The topics in Table 4.5 were similar to the topics obtained from the LDA model (Table 4.3).

Table 4.6 presents the posterior mean estimate and the corresponding 95% credibility interval for each of the parameters in the structural model component. The mean intercept estimates for Topic 1 ($\mu_{\xi_{a1}}$) and Topic 2 ($\mu_{\xi_{a2}}$) were positive. The 95% credibility interval for the $\mu_{\xi_{a1}}$ estimate did not include 0, but the 95% credibility interval for the $\mu_{\xi_{a2}}$ estimate did include 0. The two estimates indicated that Topic 1 proportions and Topic 2 proportions at time 1 tended to be higher than Topic 3 proportions at time 1. However, the Topic 2 proportions and Topic 3 proportions at time 1 could be similar since the credibility interval for $\mu_{\xi_{a2}}$ included 0. The posterior mean estimate of the slope for Topic 1 ($\mu_{\xi_{b1}}$) was negative, and the posterior mean estimate of the slope for Topic 2 ($\mu_{\xi_{b2}}$) was positive. The 95% credibility interval for the $\mu_{\xi_{b1}}$ estimate included 0, but the 95% credibility interval for the $\mu_{\xi_{b2}}$ estimate did not include 0. Thus, the two estimates indicated that the log odds of using Topic 1 rather than using Topic 3 tended to decrease in documents over time, and the log odds of using Topic 2 rather than using Topic 3 tended to increase in documents over time. In other words, as students took the test multiple times, Topic 1 proportions tended to decrease compared to the change in Topic 3 proportions,

and Topic 2 proportions tended to increase compared to the change in Topic 3 proportions. However, the change in Topic 1 proportions appear to be similar to the change in Topic 3 proportions since the credibility interval for $\mu_{\xi_{b1}}$ included 0.

The $\boldsymbol{\Psi}_1(1,1)$, $\boldsymbol{\Psi}_1(2,2)$, $\boldsymbol{\Psi}_2(1,1)$ and $\boldsymbol{\Psi}_2(2,2)$ estimates showed that the log odds of using Topic 1 and Topic 2 at time 1 varied substantially across students, and also the rate of change for the two topics varied substantially across students. Both Topic 1 and Topic 2 had negative covariances between the intercept and slope variables, which indicated that high values for the log odds of using Topic 1 and Topic 2 at time 1 were associated with the opposite direction of the corresponding topic's rate of change in the log odds. The 95% credibility intervals for $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ components did not include 0.

Bayesian Posterior Predictive Checking

A posterior predictive checking approach (Gelman, Meng, & Stern, 1996; Gelman et al., 2013) was used to evaluate whether data support the assumption of the topic modeling component. Posterior predictive checks can provide a way to evaluate where the model fits the data and where the model does not fit the data. The basic idea of this method is that replicated observations y^{rep} are simulated first from its posterior predictive distribution, which is presented as

$$p(y^{rep}|y) = \int p(y^{rep}|\boldsymbol{\Theta})p(\boldsymbol{\Theta}|y)d\boldsymbol{\Theta},$$

where $\boldsymbol{\Theta}$ is a set of model parameters. Then, these replicated observations are compared to the observed data using a function that reflects any attributes of interest. It can be set as a minimum value of y, average value of y, or some value determined by researchers depending on aspects that they want to evaluate.

Mimno and Blei (2011) used instantaneous mutual information (IMI) in their posterior predictive check for goodness of fit of LDA to evaluate the local independence assumption of LDA. LDA assumes that each word is independently drawn from its topic's multinomial distribution. Suppose that w is an observed word, d is a document index, and k is a topic index. IMI is defined as follows (Mimno & Blei, 2011):

$$IMI(w, D|k) = \sum_d P(d|w,k) \log P(d|w,k) - \sum_d P(d|k) \log P(d|k)$$

$$= \sum_d [P(d|w,k) \log P(d|w,k) - P(d|k) \log P(d|k)]$$

$$= \sum_d \left[\frac{N(d,w,k)}{N(w,k)} \log \frac{N(d,w,k)}{N(w,k)} - \frac{N(d,k)}{N(k)} \log \frac{N(d,k)}{N(k)} \right].$$

Given topic k, if word w appears in only a few documents, IMI $(w, D|k)$ will be high. If word w tends to appear across all documents, IMI $(w, D|k)$ will be low.

In this study, IMI was used to check words that did and did not satisfy the local independence assumption. For each word, IMI scores were calculated using the observed data and simulated data under the independence assumption. Figure 4.3 shows IMI scores for words in the three topics of the student response data. In each of the plots, the y-axis represents the rank of the words in the topic, and the x-axis represents IMI scores. Each bolded circle represents the observed IMI score of the word that is located below the circle. The red circles represent IMI scores from 20 replicated data. These red circles represent expected IMI values under the independence assumption. Words with lower observed IMI scores than corresponding expected IMI scores indicate that the words tended to be more uniformly distributed across documents than expected under the model assumption. Words with higher IMI scores than corresponding expected IMI scores indicate that the words tended to appear more on some specific documents

than expected under the model assumption. Most of the observed IMI scores for words in Figure 4.3 were within the range of corresponding expected IMI scores.

Conclusions from the Empirical Study

The analysis of the student response data using the growth curve topic model investigated latent topics in the corpus of students' written responses and described how topic proportions in students' responses changed over time. Three topics were detected from the students' responses. The three topics were related to everyday conversational words, discipline-specific words, and general technical words. The structural model component detected the changes in topic proportions over time. At time 1, topic proportions for Topic 1were higher than those for Topic 3 in general, and Topic 2 proportions appeared to be similar to Topic 3 proportions. As students took the test multiple times, Topic 2 proportions tended to increase compared to the change in Topic 3 proportions. The change in Topic 1 proportions were largely similar to the change in Topic 3 proportions. This tendency was also observed in the LDA analysis.

For those words that had high probability estimates for Topic 2, many were related to questions at the end of the test or technical vocabulary that was associated with scientific processes (Kim et al., 2017). The change in Topic 2 proportions may indicate either that students were more likely to complete the test as they took the test multiple times or that this change may reflect instructional effect in that they learned words related to the scientific processes. However, it should be noted that there is a limit to the interpretation of our findings because of missing responses in this data. The main purpose of this empirical data analysis was the demonstration of the growth curve topic model using a real dataset.

Simulation Study

A simulation study was conducted to evaluate the performance of the growth curve topic model with respect to the recovery of parameters in the structural model component. The focus of this simulation was to investigate the impact of sample size (the number of documents for each time point and average document length) and the variances of the mean and slope variables on parameter estimation.

Simulation Study Design

Datasets were generated using the results that were obtained from the empirical data analysis. That is, the number of topics was fixed at 3, and the word distributions for the topics were taken from the empirical data analysis. The last topic (i.e., Topic 3) was assigned as the reference topic. For each time point, three different numbers of documents (D = 150, 250, and 500) were considered, in order to be able to determine the effects of the number of documents on parameter estimation. In addition, three different average document lengths (L = 50, 100, 150) were considered to determine the impact of average document length on parameter estimation. For each of the combinations of document size and average document length, datasets were generated with different values of $\boldsymbol{\Psi}_k$, which is the variance-covariance matrix of $\boldsymbol{\xi}_{i,k} = \left(\xi_{a,i,k}, \xi_{b,i,k}\right)^T$. The correlation between $\xi_{a,i,k}$ and $\xi_{b,i,k}$ was fixed at -0.6, and two different sets of values were considered for the variances of $\xi_{a,i,k}$ and $\xi_{b,i,k}$ (these are indicated as V1 and V2) in order to examine the impact of variance size on $\mu_{\boldsymbol{\xi}_k}$ estimation. The V1 condition had relatively smaller variances in the initial status and the rate of change, and the V2 condition had greater variances in the initial status and the rate of change. $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ were set to be the same within each of the V1 and V2 conditions. The generating values of $\boldsymbol{\mu}_{\boldsymbol{\xi}_1}$ and $\boldsymbol{\mu}_{\boldsymbol{\xi}_2}$ were set to the

values that were similar to the posterior mean estimates of $\boldsymbol{\mu}_{\xi_1}$ and $\boldsymbol{\mu}_{\xi_2}$ from the empirical data analysis. The data generating conditions are summarized in Table 4.7.

Estimation

In this simulation study, the number of topics was assumed to be known as $K = 3$. The Gibbs sampling algorithm was used with the following prior distributions:

1) $\boldsymbol{\gamma}_k \sim Dirichlet(\boldsymbol{\beta})$ with $\boldsymbol{\beta} = \mathbf{1}_V$ for $k = 1,2,3$.
2) $\boldsymbol{\mu}_{\xi_k} \sim N(\boldsymbol{\mu}_{0k}, \boldsymbol{\Sigma}_{0k})$ with $\boldsymbol{\mu}_{0k} = (0,0)^T$ and $\boldsymbol{\Sigma}_{0k} = 100\boldsymbol{I}_2$ for $k = 1,2$.
3) $\boldsymbol{\Psi}_k \sim IW(v_0, \boldsymbol{S}_0)$ with $v_0 = 3$ and $\boldsymbol{S}_0 = \boldsymbol{I}_2$ for $k = 1,2$.

The tuning parameter for sampling $\boldsymbol{\xi}_{i,k}$ was set to have an acceptance rate of about 0.4. The total number of iterations was 75,000. The first 25,000 iterations were discarded for burn-in and the next 50,000 iterations were used for making posterior inferences. The post burn-in MCMC chain was thinned to every 5th iteration. That is, every 5th iteration was recorded, so the total number of recorded post burn-in iterations was 10,000. The results were summarized using 100 converged replications. Convergence was examined by Geweke's convergence diagnostic (Geweke, 1992), Heidelberger and Welch convergence diagnostic (Heidelberger & Welch, 1983), and trace plots.

Label Switching

In this simulation study, topics from the empirical data analysis were used to generate data. When estimating topics in the growth curve topic model, the order of the topics may not be the same across replications. This is referred to as label switching. This kind of label switching can be observed by comparing the generated order of the topics with the resulting model. Those replications in which the last topic (i.e., the reference topic) was either Topic 1 or Topic 2 are not appropriate for the results summary of this simulation study because this simulation study was designed to have Topic 3 as the reference topic. For example, if the last topic was Topic 2 for

some replications, then the structural model parameters would be estimated with Topic 2 as the reference topic. This would be different from the true model, which used Topic 3 as the reference topic. In order to avoid any possible label switching and have Topic 3 as the reference topic, restrictions were imposed on $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3\}$ using selected words that represents a given topic. For example, the sum of $\boldsymbol{\gamma}_3$ components that correspond to the representative words for Topic 3 was constrained to be larger than the sum of $\boldsymbol{\gamma}_2$ components that correspond to the same representative words or the sum of $\boldsymbol{\gamma}_3$ components that correspond to the same representative words. The selected representative words for a topic consisted of words that were distinguishable with regard to the probabilities assigned to the words for the three topics. That is, words that had high probabilities for a given topic but not for the other topics were selected as the representative words for that topic.

Results

For each condition, posterior mean estimates of $\boldsymbol{\mu}_{\xi_1}$, $\boldsymbol{\mu}_{\xi_2}$, $\boldsymbol{\Psi}_1$, and $\boldsymbol{\Psi}_2$ were evaluated by obtaining mean absolute bias, standard errors, and coverage rates. The mean absolute bias was defined as $\zeta^{(AB)} = \frac{1}{R}\sum_{r=1}^{R}\left|\hat{\zeta}_{(r)} - \zeta\right|$, where R was the number of replications, ζ was the generating value, and $\hat{\zeta}_{(r)}$ was a posterior mean estimate of ζ from the rth replication. The coverage rate was defined as $CvR = \frac{1}{R} Ind\left(l_{\hat{\zeta}_{(r)}} \leq \zeta \leq u_{\hat{\zeta}_{(r)}}\right)$, where $l_{\hat{\zeta}_{(r)}}$ and $u_{\hat{\zeta}_{(r)}}$ were the lower and upper bounds of the 95% credibility interval for $\hat{\zeta}_{(r)}$, respectively, and $Ind(\cdot)$ was an indicator function that had 1when ζ was within the interval and 0 otherwise.

Tables 4.8 and 4.9 summarize the results for $\boldsymbol{\mu}_{\xi_1} = \left(\mu_{\xi_{a1}}, \mu_{\xi_{b1}}\right)^T$ and $\boldsymbol{\mu}_{\xi_2} = \left(\mu_{\xi_{a2}}, \mu_{\xi_{b2}}\right)^T$ components, respectively. As shown in Table 4.8, the mean absolute bias and standard errors for $\mu_{\xi_{a1}}$ and $\mu_{\xi_{b1}}$ generally decreased as the average document length (i.e., L) increased or as the

number of documents (i.e., D) increased for both variance conditions (i.e., V1 and V2). It appeared that the number of documents had more apparent effect on decreasing mean absolute bias than average document length. Mean absolute bias values generally decreased as average document length increased, but the values were quite similar. The coverage rate for $\mu_{\xi_{a1}}$ ranged from 0.90 to 0.98 for the V1 condition and ranged from 0.90 to 97 for the V2 condition. The coverage rate for $\mu_{\xi_{b1}}$ ranged from 0.87 to 0.94 for the V1 condition and ranged from 0.89 to 0.95 for the V2 condition. As can be seen in Table 4.9, mean absolute bias and standard errors for $\mu_{\xi_{a2}}$ and $\mu_{\xi_{b2}}$ tended to decrease as the average document length increased or as the number of documents increased for both variance conditions. Similar to the results found in Table 4.8, the decreasing pattern of mean absolute bias was more apparent for the number of documents than for the average document length. The coverage rate for $\mu_{\xi_{a2}}$ ranged from 0.83 to 0.99 for the V1 condition and ranged from 0.92 to 0.98 for the V2 condition. The coverage rate for $\mu_{\xi_{b2}}$ ranged from 0.89 to 0.99 for the V1 condition and ranged from 0.92 to 0.96 for the V2 condition. For both tables, the values of absolute bias and standard errors under V2 tended to be higher than those values under V1, although the differences were not large.

Tables 4.10 and 4.11 summarize the results for $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ components, respectively. In Table 4.10, it can be seen that the mean absolute bias and standard errors for $\boldsymbol{\Psi}_1(1,1)$, $\boldsymbol{\Psi}_1(1,2)$, and $\boldsymbol{\Psi}_1(2,2)$ tended to decrease as the average document length increased or as the number of documents increased for both variance conditions. As for the coverage rate, $\boldsymbol{\Psi}_1(1,1)$, $\boldsymbol{\Psi}_1(1,2)$, and $\boldsymbol{\Psi}_1(2,2)$ ranged from 0.63 to 0.99, 0.78 to 0.97, and 0.78 to 0.96, respectively, for V1, and ranged from 0.61 to 0.98, 0.76 to 0.98, and 0.69 to 0.91, respectively, for V2. As shown in Table 4.11, the mean absolute bias and standard errors for $\boldsymbol{\Psi}_2(1,1)$, $\boldsymbol{\Psi}_2(1,2)$, and $\boldsymbol{\Psi}_2(2,2)$ generally decreased as the average document length increased or as the document size increased for both

variance conditions, which was similar to the results shown in Table 4.10. As for the coverage rate, $\boldsymbol{\Psi}_2(1,1)$, $\boldsymbol{\Psi}_2(1,2)$, and $\boldsymbol{\Psi}_2(2,2)$ ranged from 0.83 to 0.97, 0.77 to 0.94, and 0.83 to 0.98, respectively, for V1, and ranged from 0.78 to 0.95, 0.81 to 0.95, and 0.80 to 0.95, respectively, for V2. The coverage rate tended to decrease as average document length or document size increased for both tables. Further investigation of credibility interval widths indicated that the widths generally decreased considerably compared to the reduction in the size of bias. This decrease led to decreasing coverage rates when average document length or document size increased. In addition, as can be seen in both tables, absolute bias values and standard errors under the V2 condition tended to be higher than those values under the V1 condition.

Simulation Study Conclusions

The results suggest that both average document length and the number of documents influence parameter recovery. In general, the longer the average document length, the smaller the mean absolute bias values, and the larger the number of documents, the smaller the mean absolute bias values. However, the effect of the number of documents was more apparent than that of the average document length on parameter recovery. Standard errors also decreased in general as average document length or the number of documents increased. As the generating variance increased (i.e., from V1 to V2), the mean absolute bias and standard errors tended to increase. The coverages rates for $\boldsymbol{\mu}_{\xi_1}$ and $\boldsymbol{\mu}_{\xi_2}$ components did not appear to have consistent patterns, but the coverage rates for $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ components appeared to decrease when average document length or the number of documents increased.

Discussion

In this study, the growth curve topic model was introduced for analyzing text data collected over multiple time points. The model was designed to capture latent topics in texts and

then to help characterize the change in topic proportions. An example of the practical use of this model in an educational setting was illustrated using empirical data from an NSF-funded host study. A simulation study was also presented to investigate the impact of sample size and the variances of mean and slope variables on parameter estimation.

In the empirical data analysis, topics in students' responses were detected and the change in students' use of topics was examined over four time points. The results indicated that certain patterns of changes in topic proportions were present in the longitudinal data. These changes were consistent with the changes in topic proportions detected by LDA. Results from the empirical data analysis were also consistent with findings from previous research that used other albeit similar data from the same host study. In this regard, Buxton et al. (2014) reported results from a qualitative analysis that showed an increased use of technical vocabulary and an increased percentage of lexical density in the posttest responses compared to the pretest responses. Technical vocabulary and lexical density are types of linguistic features. Technical vocabulary includes the appropriate use of scientific words. Lexical density is measured by the ratio of content words to grammatical words. Typically, academic writing has higher lexical density than non-technical writing. Kim et al. (2017) found that students generally used more words from a topic characterized as discipline specific language on the posttest. This pattern was also found in the qualitative analysis in the form of increased use of technical vocabulary and lexical density. The current study with the growth curve topic model effectively captured these changes in students' responses using a linear growth curve model.

The simulation study results indicated that the bias of parameter estimates tended to decrease as the average document length increased or as the number of documents increased. However, the number of documents appeared to have more effect on parameter estimation than

average document length. The mean absolute values for the shortest average document length (50) and the longest average document length (150) did not appear to differ appreciably. In addition, parameter estimates, particularly variance components, from conditions having larger variances of mean and slope variables (i.e., V2) appeared to have larger mean absolute bias values. These results suggest that having documents with 50 words on average would be sufficient for the growth curve topic model analysis, and that the bias of parameter estimates can be further controlled by the number of documents. In addition, collecting a large number of documents would be recommended, when large variances of mean and slope variables are expected.

In this study, topics in the growth curve topic model were fixed across all time points. This model can be extended by allowing topics to evolve over time, as is possible with the dynamic topic model (Blei & Lafferty, 2006). If we allow topics to change over time, topics from one time point will be different from topics from another time point, and the interpretation of structural model parameters, especially $\boldsymbol{\mu}_{\xi_1}$ and $\boldsymbol{\mu}_{\xi_2}$, might be different from the current study.

We conclude this study by specifying some areas that we want to explore for future research. Roberts et al. (2016) introduced a topic model that incorporates covariates to investigate the relationship between covariates and topics. The growth curve topic model can also be extended to include covariates to predict changes in topic proportions. Variables such as gender or type of intervention may provide useful information on change in topic proportions. In addition, text data in education are sometimes associated with scores. Some topic models jointly models texts and associated scores. For example, supervised topic models (e.g., Blei & McAuliffe, 2008) deal with both documents and associated scores. Extending the growth curve topic model for jointly model score variables is another possible direction for future research.

References

Attali, Y. (2014). A ranking method for evaluating constructed responses. *Educational and Psychological Measurement*, *74*(5), 795-808.

Bennett, R. E. (1991). On the meanings of constructed response. *ETS Research Report Series*, *1991*, i-46, doi:10.1002/j.2333-8504.1991.tb01429.x

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (pp. 113-120). ACM.

Blei, D. M., & McAuliffe, J. D. (2008). Supervised topic models. In *Advances in neural information processing systems* (pp. 121-128).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*, 993-1022.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective* (Vol. 467). John Wiley & Sons.

Buxton, C., Allexsaht-Snider, M., Aghasaleh, R., Kayumova, S., Kim, S., Choi, Y., & Cohen, A. (2014). Potential benefits of bilingual constructed responses science assessments for emergent bilingual learners. *Double Helix*, *2*(1), 1-21.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J.D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22* (pp. 288-296). Cambridge, MA: The MIT Press.

Chen, Y., Yu, B., Zhang, X., & Yu, Y. (2016). Topic modeling for evaluating students' reflective writing: a case study of pre-service teachers' journals. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 1-5). ACM.

Chib, S., & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The american statistician*, *49*(4), 327-335.

Cillessen, A. H., & Mayeux, L. (2004). From censure to reinforcement: Developmental changes in the association between aggression and social status. *Child development*, *75*(1), 147-163.

Crossley, S., Salsbury, T., & McNamara, D. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, *59*(2), 307-334.

Crossley, S. A., Salsbury, T., & Mcnamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, *36*(5), 570-590.

Curran, P. J., Harford, T. C., & Muthen, B. O. (1996). The relation between heavy alcohol use and bar patronage: a latent growth model. *Journal of studies on alcohol*, *57*(4), 410-418.

Curran, P. J., & Muthén, B. O. (1999). The application of latent curve analysis to testing developmental theories in intervention research. *American Journal of Community Psychology*, *27*(4), 567-595.

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development*, *11*(2), 121-136.

Daller, M., Turlik, J., & Weir, I. (2013). Vocabulary acquisition and the learning curve. In S. Jarvis & M. Daller (Eds.), *Vocabulary Knowledge: Human ratings and automated measures* (Vol. 47, pp. 185-218). Amsterdam, John Benjamins.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford.

Furtak, E. M., & Ruiz-Primo, M. A. (2008). Making students' thinking explicit in writing and discussion: An analysis of formative assessment prompts. *Science Education*, *92*(5), 799-824.

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, *85*(410), 398-409.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman and Hall/CRC.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, 733-760.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6), 721-741.

Gerrish, S., & Blei, D. M. (2010). A language-based approach to measuring scholarly impact. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (pp. 375-382).

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (with discussion). In J. O., Berger, J. M. Bernardo, A. P. Dawid, & A. F. M. Smith (Eds.) *Bayesian Statistics 4* (pp. 156-163). Oxford: Oxford University Press.

Glynn, C., Tokdar, S. T., Howard, B., & Banks, D. L. (2019). Bayesian Analysis of Dynamic Linear Topic Models. *Bayesian Analysis*, *14*(1), 53-80.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228-5235.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267-297.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97-109.

Heidelberger, P., & Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, *31*(6), 1109-1144.

Hofmann, T. (1999, July). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (pp. 289-296). Morgan Kaufmann Publishers Inc.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, *42*(1-2), 177-196.

Johri, A., Wang, G. A., Liu, X., & Madhavan, K. (2011). Utilizing topic modeling techniques to identify the emergence and growth of research topics in engineering education. In *2011 Frontiers in Education Conference (FIE)*, (pp. T2F-1-T2F-6). Rapid City, SD. doi: 10.1109/FIE.2011.6142770

Kim, S., Kwak, M., Cardozo-Gaibisso, L., Buxton, C., & Cohen, A. S. (2017). Statistical and Qualitative Analyses of Students' Answers To a Constructed Response Test of Science Inquiry Knowledge. *Journal of Writing Analytics*, *1*, 82-102.

Lauderdale, B. E., & Clark, T. S. (2014). Scaling politically meaningful dimensions using texts and votes. *American Journal of Political Science*, *58*(3), 754-771.

Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, *5*(1), 1608.

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis, 23*(2), 254-277.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual review of psychology*, *60*, 577-605.

McCoach, D. B., O'Connell, A. A., Reis, S. M., & Levitt, H. A. (2006). Growing readers: A hierarchical linear model of children's reading growth during the first 2 years of school. *Journal of Educational Psychology*, *98*(1), 14-28.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, *21*(6), 1087-1092.

Mimno, D., & Blei, D. (2011). Bayesian checking for topic models. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 227-237). Association for Computational Linguistics.

Muthén, B. O., & Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological methods*, *2*(4), 371-402.

Owens, E. B., & Shaw, D. S. (2003). Predicting growth curves of externalizing behavior across the preschool years. *Journal of abnormal child psychology*, *31*(6), 575-590.

Paul, M. J., & Dredze, M. (2011). You are what you tweet: analyzing Twitter for public health. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media* (pp. 265-272).

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R news, 6*(1), 7-11.

Raftery, A. E., Newton, M. A., Satagopan, J. M., & Krivitsky, P. N. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity (with Discussion) In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, & M. West (Eds.) *Bayesian Statistics 8*. Oxford: Oxford University Press.

Reich, J., Tingley, D., Leder-Luis, J., Roberts, M. E., & Stewart, B. M. (2015). Computer-assisted reading and discovery for student-generated text in massive open online courses. *Journal of Learning Analytics*, *2*(1), 156-184.

Rhemtulla, M., & Hancock, G. R. (2016). Planned missing data designs in educational psychology research. *Educational Psychologist*, *51*(3-4), 305-316.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B, & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*(4), 1064-1082.

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, *111*(515), 988-1003.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, *40*(2), 163-184.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological bulletin*, *92*(3), 726-748.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods, 7*(2), 147-177.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of educational and behavioral statistics*, *23*(4), 323-355.

Socher, R., Gershman, S., Sederberg, P., Norman, K., Perotte, A. J., & Blei, D. M. (2009). A Bayesian analysis of dynamics in free recall. In Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems* 22 (pp. 1714-1722). Cambridge, MA: MIT.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583-639.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, *82*(398), 528-540.

Van Dyk, D. A., & Meng, X. L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, *10*(1), 1-50.

Wang, C., Blei, D., & Heckerman, D. (2012). Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298*.

Wang, C., Song, T., Wang, Z., & Wolfe, E. (2017). Essay selection methods for adaptive rater monitoring. *Applied Psychological Measurement, 41*(1), 60-79.

Williams, S. T., Conger, K. J., & Blozis, S. A. (2007). The development of interpersonal aggression during adolescence: The importance of parents, siblings, and family economics. *Child Development, 78*(5), 1526-1542.

CHAPTER 5

CONCLUDING REMARKS AND FUTURE STUDIES

In education, constructed response items are often used to understand students' cognitive processes that multiple-choice items may not reveal. Constructed response items can provide data that consist of students' written responses and scored data based on their written responses. This dissertation addressed different approaches to modeling data obtained from constructed response items. The dissertation first discussed estimating reliability for categorical responses. This was initially motivated by an attempt to accurately estimate internal consistency reliability for a test consisting of constructed response items with items scored with different numbers of ordered categories. The second and third studies focused on students' written response to constructed response items and proposed new models for extracting information from the responses. This final chapter summarizes findings from the three studies and discusses potential future research.

Study 1: Reliability for Tests with Items Having Different Numbers of Ordered Categories

There are many methods to estimate the reliability of a test, and structural equation modeling has been shown to be one that is useful in this regard. In this study, we focused on estimating reliability for categorical data. Constructed response items are generally scored with categorical values. The range of the categorical values depends on a rubric and often is not the same across the items. In this study, we used the structural equation modeling approach to estimate the reliability of a test consisting of items having unequal numbers of ordered categories. The proposed nonlinear SEM reliability can handle situations in which a test has

items with different numbers of ordered categories and structures such as testlets. We provided a numerical calculation to derive the nonlinear SEM reliability and presented a simulation and an empirical data analysis for evaluating the performance of the nonlinear SEM reliability.

Results from the simulation study showed that the nonlinear SEM reliability estimate was close to the population reliability across all conditions. The nonlinear SEM reliability estimate and population reliability had their lowest values when the number of categories was two and had their largest values when the number of categories was five. The reliability values for conditions with a mixed number of categories were located in-between. For purposes of comparison, coefficient alpha was compared with the nonlinear SEM reliability. Coefficient alpha was close to the population reliability for simulation conditions under a one-factor model and for equal numbers of categories. Coefficient alpha values were lower than the population reliability, however, for conditions in which the data fit a bi-factor model or had items with unequal numbers of response categories.

In the empirical data analysis, a one-factor model and a bi-factor model were fit to data from a test consisting of 27 constructed response items, and the nonlinear SEM reliability was calculated based on parameter estimates from each of the models. The items had 2 or 3 response categories and had a testlet structure. The one-factor solution did not provide a good fit and provided an unacceptable reliability coefficient value, which was larger than 1. The bi-factor solution had a better model fit and was consistent with the construction of the test. For this solution, the nonlinear SEM reliability estimate was .96. For comparison, coefficient alpha was also obtained, and the value was .92, i.e., it was lower than the nonlinear SEM reliability under the bi-factor model.

Study 2: Exploring Students' Responses to Constructed Response Items with a Supervised Mixture Topic Model for Heterogeneous Populations

This study addressed finding relationships between students' responses to constructed response items and associated scores. In general, responses to constructed response items are often paired with scores because the responses are usually graded by raters or by machine. Investigating the relationship between the responses and scores is informative, especially in the educational context, because it helps in understanding how students receive scores based on their written responses. This study also addressed situations in which the population consists of two or more latent classes, each of which is assumed to have unique relationships between labels (e.g., scores) and topic proportions. A new model, MixSLDA was introduced in this study to investigate relationships in the paired data. This model was designed to detect meaningful homogeneous subgroups regarding the relationships between topic proportions in examinees' responses and their scores on their responses. The proposed model was demonstrated using a real data analysis from an NSF-funded host study on instructing teachers on teaching understanding of the science inquiry process to middle school students. In addition, a simulation study was conducted to evaluate the performance of the MixSLDA under practical testing conditions.

In the empirical data analysis, students' responses to constructed response items and associated scores were analyzed using MixSLDA. Four topics were detected in students' written responses, and two latent classes were discovered in the relationship between students' responses and associated scores. The results showed that the relationship between students' use of the four topics and their scores differed for the two latent classes. Specifically, the relationship between scores and the topic related to the understanding of controlling variables differed for the two latent classes. The expected score for students in Class 1 by using words from the controlling

variable related topic was lower than the expected score for students in Class 2. This suggests that students belonging to Class 1 tended to fail in appropriately discussing the controlling variable related questions.

In the simulation study, the MixSLDA parameter recovery was investigated by considering topic structure, document size, and document length. Data were generated to simulate the empirical data. The results from the simulation study suggested that parameter recovery was influenced by all three components. The mean relative bias values were smaller for conditions with topics with less similarity, and the mean absolute bias and mean relative bias generally decreased as document size or document length increased. This simulation study suggested that documents having more than 50 words on average are necessary for the MixSLDA analysis given other conditions covered in this simulation study.

Study 3: Topic Modeling for Longitudinal Text Data

Many assessments include essay items or constructed response items in order to measure growth in students' in-depth knowledge and in aspects of students' performance. When text data are collected longitudinally over multiple time points, the change pattern can be an important focus of research, similar to growth curve modeling in the social and behavioral sciences. However, typical methods to model change have focused almost entirely on models and analysis of data consisting of numeric responses. For textual data, these responses are usually graded first with numeric scores, and then the scores are analyzed. In this study, we proposed the growth curve topic model, in which we incorporate topic modeling and growth curve modeling to better understand how students change their use of language in response to instruction as reflected in constructed response data. The growth curve topic model was applied to middle-grade students'

written responses to constructed response items collected over four time points. Additionally, a simulation study was presented to evaluate the performance of the growth curve topic model.

In order to illustrate the use of the growth curve topic model, an empirical data analysis using middle school students' responses to constructed response items was provided. The growth curve topic model captured three topics in students' responses and characterized changes in topic proportions over time. The general technical vocabulary-related topic was the reference topic and the other two topics were compared with the reference topic. While the use of the everyday language-related topic tended to decrease over time compared to the reference topic, the discipline specific language-related topic tended to increase over time compared to the reference topic. These results were consistent with findings from previous research that used other, albeit similar, data from the same host study.

In the simulation study, the performance of the growth curve topic model was evaluated with respect to parameter recovery of the structural model component. The impacts of document size, document length, and the variances of the mean and slope variables on parameter estimation were investigated. The results indicated that the three components all influenced parameter estimation. A larger number of documents or documents with longer length provided less biased parameter estimates. Also, parameter estimates for conditions with smaller variances tended to be less biased.

Future Directions

This dissertation consisted of three studies, each of which addressed research questions related to analyzing constructed response data. Each study has unique suggestions for future studies. First, Study 1 is based on the assumption that models are correctly specified. As the empirical data analysis indicated, nonlinear SEM reliability may fail to estimate reliability if a

wrong model is assumed. Therefore, further study is needed to illustrate the impact of model misspecification on nonlinear SEM reliability. Study 2 was conducted based on continuous scores. In education, however, essay responses are often paired with categorical scores. Extensions of MixSLDA to consider categorical types of scores would be useful for expanding the application of this model. Finally, with respect to Study 3, the growth curve topic model can be extended to include covariates, such as grade level or teachers' participation in the project, to explain the change in topic proportions. Additionally, the current growth curve topic model explored text data only. If text data are associated with scores, incorporating score information into the growth curve topic model would be useful for understanding the interactions between scores and topics over time.

Table 2.1

Factor Structures Considered in the Simulation Study

	Number of items	Number of group factors	Number of items per factor
Model 1	9	NA	NA
Model 2	18	NA	NA
Model 3	9	3	3
Model 4	18	3	6

Table 2.2

The Number of Converged Replications (Model 3 and Model 4)

		Normal			Moderate skewness			Mixed skewness		
Model	Gr	C2	C5	C25	C2	C5	C25	C2	C5	C25
3	.4	89	99	95	76	100	87	41	100	78
3	.6	100	100	99	96	100	100	94	100	92
4	.4	100	100	100	100	100	100	98	100	100
4	.6	100	100	100	100	100	100	82	100	100

Note. Gr refers to Group factor loading, and C refers to the condition for the number of response categories

Table 2.3

The Means of Reliability Coefficients and their Standard Deviations for Data from the One-factor Models

			Normal			Moderate skewness			Mixed skewness		
M	Gr		C2	C5	C25	C2	C5	C25	C2	C5	C25
1	NA	$\rho_{X\tilde{X}}$	0.81	0.87	0.83	0.79	0.87	0.81	0.77	0.86	0.80
		ρ_{non}	0.82	0.88	0.84	0.80	0.87	0.82	0.79	0.86	0.80
			(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
		α	0.81	0.87	0.82	0.79	0.87	0.78	0.76	0.85	0.75
			(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
2	NA	$\rho_{X\tilde{X}}$	0.90	0.93	0.91	0.88	0.93	0.90	0.87	0.92	0.89
		ρ_{non}	0.90	0.94	0.92	0.89	0.93	0.90	0.88	0.93	0.89
			(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)
		α	0.90	0.93	0.90	0.88	0.93	0.88	0.87	0.92	0.87
			(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)

Note. M and Gr refer to Model and Group factor loading, respectively, C refers to the condition for the number of response categories, and $\rho_{X\tilde{X}}$, ρ_{non}, and α refer to the population reliability for observed sum scores, nonlinear SEM reliability, and coefficient alpha, respectively.

Table 2.4

The Means of Reliability Coefficients and their Standard Deviations for Data from the Bifactor Models

			Normal			Moderate skewness			Mixed skewness		
M	Gr		C2	C5	C25	C2	C5	C25	C2	C5	C25
3	0.4	$\rho_{X\tilde{X}}$	0.86	0.91	0.88	0.84	0.91	0.87	0.82	0.90	0.85
		ρ_{non}	0.86	0.91	0.89	0.85	0.91	0.87	0.84	0.90	0.86
			(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
		α	0.83	0.89	0.84	0.81	0.88	0.80	0.78	0.86	0.77
			(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
3	0.6	$\rho_{X\tilde{X}}$	0.92	0.95	0.93	0.91	0.96	0.93	0.89	0.95	0.92
		ρ_{non}	0.92	0.95	0.93	0.91	0.96	0.93	0.89	0.95	0.93
			(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)
		α	0.86	0.90	0.85	0.85	0.90	0.83	0.79	0.88	0.78
			(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
4	0.4	$\rho_{X\tilde{X}}$	0.92	0.95	0.94	0.91	0.95	0.93	0.90	0.95	0.92
		ρ_{non}	0.93	0.96	0.94	0.92	0.95	0.93	0.91	0.95	0.93
			(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)
		α	0.91	0.94	0.91	0.90	0.94	0.90	0.88	0.93	0.88
			(0.01)	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)
4	0.6	$\rho_{X\tilde{X}}$	0.96	0.98	0.97	0.95	0.98	0.96	0.94	0.97	0.96
		ρ_{non}	0.96	0.98	0.97	0.96	0.98	0.97	0.95	0.98	0.96
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
		α	0.93	0.95	0.93	0.92	0.95	0.91	0.90	0.94	0.89
			(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)

Note. M and Gr refer to Model and Group factor loading, respectively, C refers to the condition for the number of response categories, and $\rho_{X\tilde{X}}$, ρ_{non}, and α refer to the population reliability for observed sum scores, nonlinear SEM reliability, and coefficient alpha, respectively.

Table 2.5

Descriptive Statistics for Each Item in the Assessment

Item	C	M	SD	Skewness	Item	C	M	SD	Skewness
1	2	0.56	0.50	-0.26	15	3	0.50	0.76	1.13
2	2	0.55	0.50	-0.20	16	3	0.73	0.64	0.32
3	2	0.60	0.49	-0.41	17	2	0.32	0.47	0.76
4	2	0.54	0.50	-0.16	18	2	0.24	0.43	1.22
5	3	1.10	0.87	-0.20	19	3	0.64	0.82	0.74
6	2	0.71	0.45	-0.92	20	3	0.77	0.68	0.31
7	2	0.51	0.50	-0.04	21	2	0.63	0.48	-0.55
8	2	0.68	0.47	-0.76	22	3	0.63	0.87	0.78
9	2	0.36	0.48	0.58	23	3	0.57	0.61	0.55
10	2	0.43	0.50	0.28	24	2	0.59	0.49	-0.35
11	2	0.45	0.50	0.19	25	3	0.57	0.84	0.94
12	3	0.58	0.65	0.67	26	2	0.18	0.38	1.68
13	2	0.11	0.31	2.52	27	3	0.42	0.59	1.07
14	2	0.24	0.42	1.25					

Note. C refers to the number of categories, and *M* and SD refer to the mean score and the standard deviation of the score, respectively.

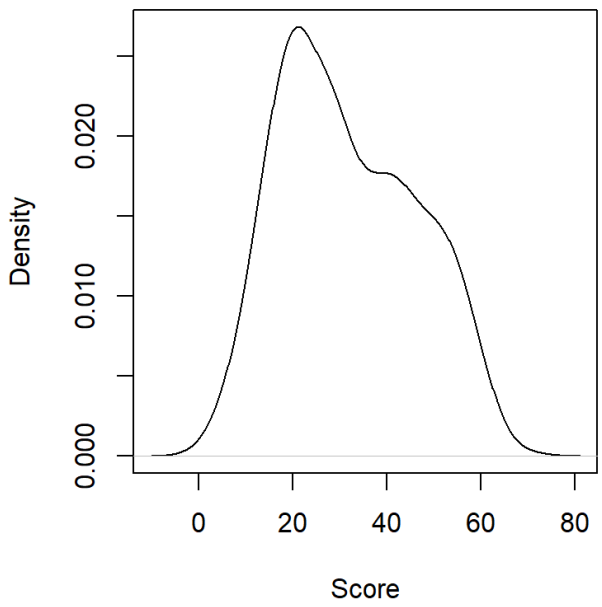

Figure 3.1. The density plot of the scores.

Table 3.1

DIC and AICM Values for LDA Models with Two to Seven Topics.

	DIC	AICM
2	476249	-476780
3	473659	-474856
4	472688	**-474603**
5	471973	-474950
6	471499	-475590
7	**471221**	-476134

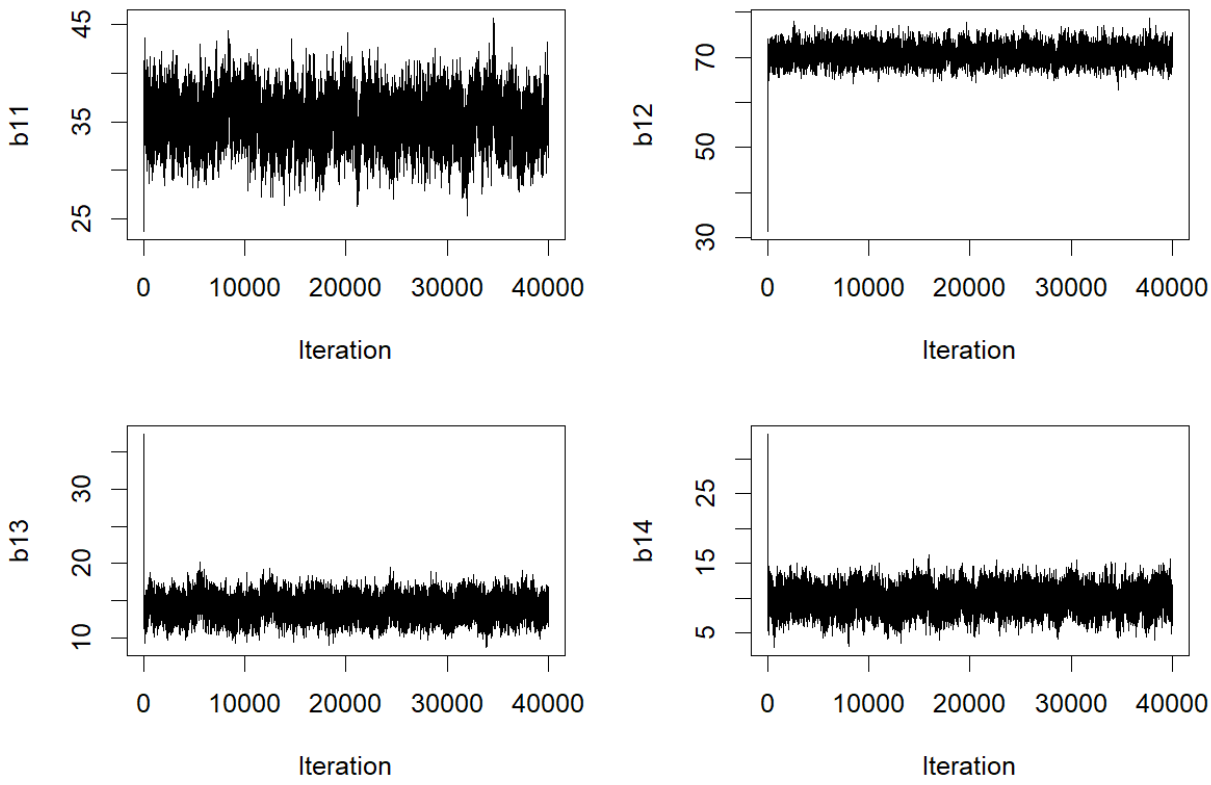

Figure 3.2. Trace plots of $\boldsymbol{b} = (b_{11}, b_{12}, b_{13}, b_{14})^T$.

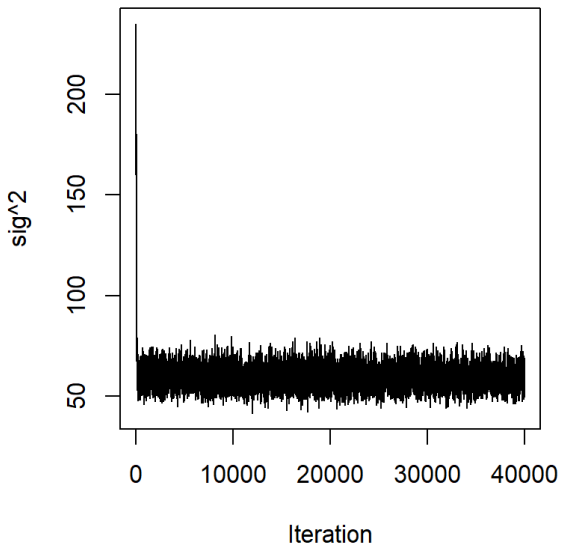

Figure 3.3. Trace plot of σ^2.

Table 3.2

Top 18 Words Having the Highest Posterior Means of Probabilities

	Topic 1		Topic 2		Topic 3		Topic 4	
	Word	Prop.	Word	Prop.	Word	Prop.	Word	Prop.
1	clone	0.049	salt	0.032	fish	0.075	fish	0.045
2	sheep	0.048	temperature	0.030	algae	0.058	because	0.038
3	fish	0.029	water	0.024	small	0.043	more	0.026
4	weight	0.029	same	0.023	cause	0.041	if	0.024
5	non	0.022	increase	0.021	effect	0.040	make	0.024
6	sick	0.022	boil	0.020	population	0.034	eat	0.023
7	people	0.021	energy	0.020	eat	0.032	water	0.020
8	because	0.021	algae	0.019	increase	0.029	salt	0.018
9	time	0.021	if	0.019	because	0.029	die	0.017
10	type	0.020	because	0.019	energy	0.021	one	0.016
11	energy	0.017	faster	0.017	converter	0.021	honey	0.015
12	small	0.014	more	0.016	catalytic	0.018	sick	0.015
13	if	0.014	sheep	0.016	disease	0.018	big	0.014
14	lifting	0.014	pour	0.014	decrease	0.016	much	0.013
15	catalytic	0.014	viscosity	0.014	die	0.015	hot	0.012
16	algae	0.013	hotter	0.014	work	0.013	what	0.012
17	converter	0.013	honey	0.014	kill	0.013	out	0.012
18	temperature	0.013	hypothesis	0.014	large	0.013	up	0.011

Table 3.3

Averages of Posterior Means of Topic Proportions and the Posterior Means of the Coefficient Parameters

Topic	$\boldsymbol{\theta}$	$\boldsymbol{b}$
1	0.23	35.09
		(30.55, 39.56)
2	0.25	70.72
		(67.45, 74.12)
3	0.27	14.04
		(11.53, 16.58)
4	0.25	9.88
		(6.86, 12.85)

Note. The values in the parentheses for each row presents the 95% credibility interval of the corresponding coefficient.

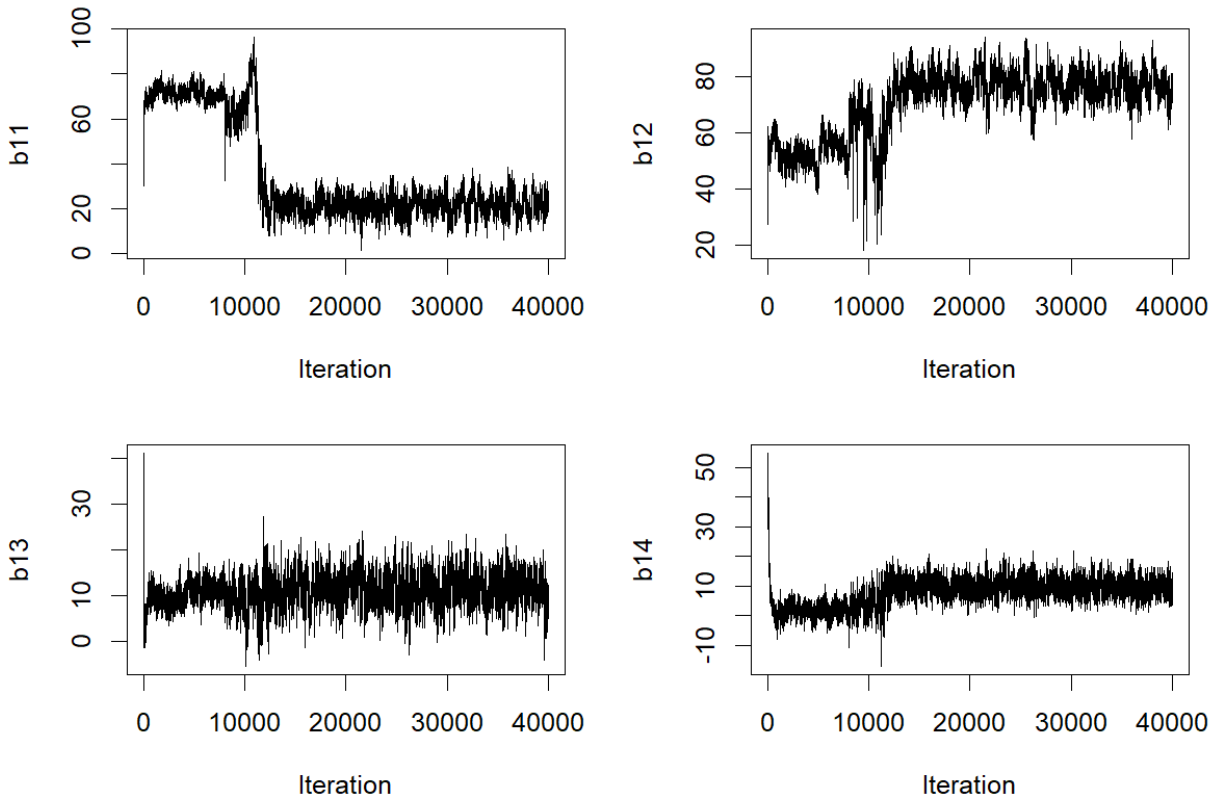

Figure 3.4. Trace plots of the components of $\boldsymbol{b}_1 = (b_{11}, b_{12}, b_{13}, b_{14})^T$.

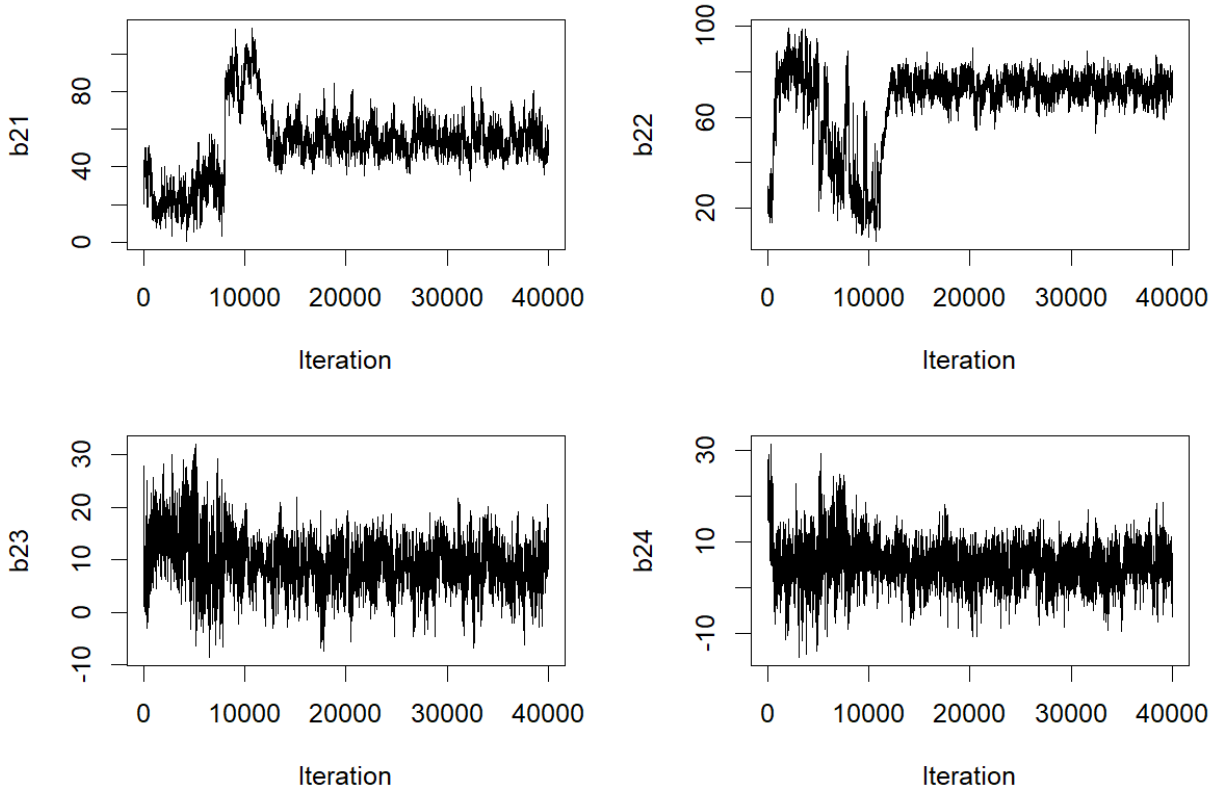

Figure 3.5. Trace plots of the components of $\boldsymbol{b}_2 = (b_{21}, b_{22}, b_{23}, b_{24})^T$.

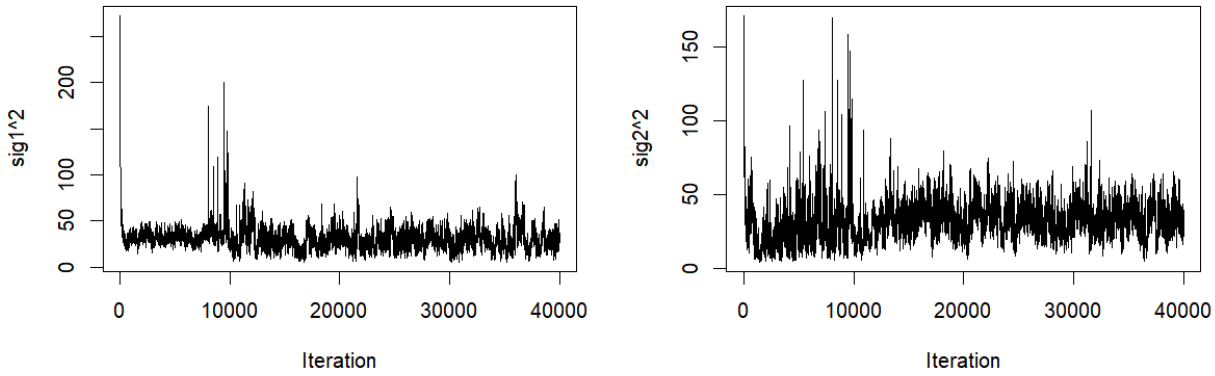

Figure 3.6. Trace plots of σ_1^2 (left side) and σ_2^2 (right side).

Table 3.4

Top 18 Words Having the Highest Posterior Means of Probabilities

	Topic 1		Topic 2		Topic 3		Topic 4	
	Word	Prop	Word	Prop	Word	Prop	Word	Prop
1	clone	0.048	salt	0.032	fish	0.076	fish	0.045
2	sheep	0.047	temperature	0.030	algae	0.059	because	0.038
3	fish	0.029	water	0.024	small	0.043	more	0.026
4	weight	0.028	same	0.023	cause	0.041	make	0.024
5	sick	0.022	increase	0.021	effect	0.040	eat	0.023
6	non	0.022	energy	0.020	population	0.034	if	0.023
7	people	0.021	algae	0.020	eat	0.032	water	0.019
8	time	0.021	boil	0.020	increase	0.029	die	0.018
9	because	0.020	because	0.019	because	0.029	salt	0.018
10	type	0.019	if	0.018	converter	0.021	one	0.016
11	energy	0.017	faster	0.017	energy	0.021	sick	0.015
12	if	0.015	more	0.016	catalytic	0.019	honey	0.015
13	small	0.014	sheep	0.015	disease	0.018	big	0.014
14	lifting	0.014	hotter	0.014	decrease	0.016	much	0.013
15	catalytic	0.013	pour	0.014	die	0.015	what	0.012
16	temperature	0.013	hypothesis	0.014	work	0.013	hot	0.012
17	increase	0.013	viscosity	0.014	large	0.013	out	0.012
18	converter	0.013	honey	0.013	kill	0.013	up	0.011

Table 3.5

Averages of Posterior Means of Topic Proportions and the Posterior Means of the Coefficient Parameters

	$\boldsymbol{\theta}$		$\boldsymbol{b}$	
Topics	Class 1	Class 2	Class 1	Class 2
1	0.22	0.24	21.48	54.51
			(12.77, 30.72)	(42.95, 69.41)
2	0.21	0.30	76.71	73.06
			(66.83, 86.78)	(64.36, 80.64)
3	0.28	0.25	11.47	8.60
			(4.76, 18.00)	(0.86, 15.33)
4	0.29	0.20	9.98	4.73
			(4.93, 15.63)	(-2.31, 10.80)

Note. The values in the parentheses for each row presents the 95% credibility interval of the corresponding coefficient.

Table 3.6

DIC and AICM Values for the MixSLDA Models

	DIC	AICM
1	476484	-478938
2	**476365**	**-478873**
3	476389	-478948

Table 3.7

Mean Absolute Bias and Relative Bias of $\boldsymbol{b}_1 = (b_{11}, b_{12}, b_{13}, b_{14})^T$ Estimates Under the S1 Condition

			AB				RB			
Δ	D	L	b_{11}	b_{12}	b_{13}	b_{14}	b_{11}	b_{12}	b_{13}	b_{14}
2.5	300	50	5.04	9.41	5.24	4.62	-17.28	12.15	-2.94	4.84
		100	2.97	6.16	3.37	3.42	-7.64	8.08	-12.10	5.25
		200	2.17	3.90	2.59	2.88	-2.82	4.50	-6.37	0.81
	500	50	3.26	8.29	3.77	3.66	-12.21	10.40	-17.62	1.92
		100	2.20	4.90	2.32	2.54	-4.46	6.62	-10.86	-1.25
		200	1.68	3.43	2.17	2.36	-0.94	3.86	-7.69	1.87
	1000	50	1.97	5.56	2.22	2.24	-7.17	7.70	-12.00	-4.52
		100	1.49	3.81	1.84	1.58	-3.43	5.36	-10.78	-2.02
		200	1.38	2.57	1.69	1.51	-2.10	3.18	-6.40	0.42
3.5	300	50	3.04	8.16	3.29	3.57	-6.84	9.34	-15.58	3.64
		100	2.17	4.87	2.60	2.54	-0.98	5.89	-10.22	-2.64
		200	1.47	2.98	1.69	2.17	-1.73	2.97	-4.02	-0.78
	500	50	3.11	6.48	3.17	2.49	-11.48	8.33	-5.11	1.90
		100	1.52	3.97	1.85	1.81	-1.55	5.19	-8.51	-2.26
		200	1.30	2.41	1.48	1.51	-1.72	3.01	-5.84	1.06
	1000	50	1.55	5.85	1.96	1.77	-5.05	8.09	-10.03	-3.86
		100	1.06	3.09	1.35	1.32	-2.48	4.37	-6.56	-2.87
		200	0.98	1.89	1.02	1.01	-1.79	2.35	-3.89	-0.54

Note. AB = mean absolute bias; RB = mean relative bias; $\Delta =$ mode separation; D = number of documents; L = average document length.

Table 3.8

Mean Absolute Bias and Relative Bias of $\boldsymbol{b}_2 = (b_{21}, b_{22}, b_{23}, b_{24})^T$ *Estimates Under the S1 Condition*

			AB				RB			
Δ	D	L	b_{21}	b_{22}	b_{23}	b_{24}	b_{21}	b_{22}	b_{23}	b_{24}
2.5	300	50	6.42	3.85	8.37	8.57	10.46	1.15	-16.45	-33.13
		100	3.36	2.95	3.41	4.77	4.09	0.92	-0.98	-18.07
		200	2.30	2.86	2.37	3.54	1.82	0.26	-1.40	-9.02
	500	50	3.50	3.61	3.23	6.82	4.97	3.80	-2.53	-30.92
		100	2.51	3.04	2.38	3.39	3.31	1.62	-1.31	-14.82
		200	2.00	2.27	2.08	2.64	1.85	0.11	-0.20	-8.89
	1000	50	2.42	3.74	2.19	4.49	3.75	5.15	-3.24	-21.49
		100	1.87	1.87	1.66	2.98	2.75	1.70	-1.01	-12.31
		200	1.34	1.48	1.52	1.86	1.43	0.79	0.05	-7.48
3.5	300	50	4.23	3.12	5.12	7.51	5.42	1.16	-3.91	-25.36
		100	2.54	2.93	2.56	4.01	3.36	1.65	-0.68	-15.70
		200	1.60	1.90	1.61	2.31	1.05	0.82	-0.18	-6.08
	500	50	3.06	2.80	3.35	6.65	4.64	3.31	-1.09	-29.65
		100	1.85	2.08	1.79	2.86	2.43	1.75	-0.52	-11.54
		200	1.18	1.45	1.38	1.86	1.36	0.56	-0.18	-5.34
	1000	50	2.13	3.00	1.58	3.99	3.45	4.17	-1.35	-19.39
		100	1.43	1.86	1.17	2.51	2.26	2.39	-1.10	-11.48
		200	0.93	1.02	0.99	1.34	0.71	0.91	0.09	-5.11

Note. AB = mean absolute bias; RB = mean relative bias; $\Delta =$ mode separation; D = number of documents; L = average document length.

Table 3.9

Standard Errors and Coverage Rates of $\boldsymbol{b}_1 = (b_{11}, b_{12}, b_{13}, b_{14})^T$ *Estimates Under the S1 Condition*

			SE				CP			
Δ	D	L	b_{11}	b_{12}	b_{13}	b_{14}	b_{11}	b_{12}	b_{13}	b_{14}
2.5	300	50	5.35	6.08	7.15	5.67	0.90	0.87	0.95	0.98
		100	3.32	4.60	3.66	4.17	0.96	0.86	0.96	0.96
		200	2.75	3.93	3.17	3.65	0.94	0.88	0.91	0.95
	500	50	3.23	5.70	3.59	4.32	0.91	0.70	0.85	0.95
		100	2.66	3.65	2.37	3.19	0.94	0.76	0.93	0.94
		200	2.18	3.16	2.44	2.84	0.93	0.88	0.92	0.95
	1000	50	2.07	3.29	2.11	2.65	0.90	0.68	0.90	0.95
		100	1.78	2.43	1.65	1.92	0.92	0.72	0.88	0.95
		200	1.66	2.28	1.82	2.02	0.95	0.83	0.92	0.95
3.5	300	50	3.45	6.68	3.31	4.31	0.99	0.87	0.89	0.98
		100	2.72	4.41	2.79	3.11	0.96	0.83	0.94	0.94
		200	1.90	3.37	2.22	2.73	0.96	0.90	0.92	0.94
	500	50	3.56	4.68	5.97	3.25	0.86	0.72	0.89	0.93
		100	1.84	2.99	1.89	2.22	0.96	0.79	0.90	0.96
		200	1.63	2.12	1.68	1.83	0.95	0.87	0.94	0.96
	1000	50	1.63	3.18	1.75	2.07	0.91	0.54	0.88	0.92
		100	1.24	1.99	1.31	1.59	0.96	0.74	0.92	0.92
		200	1.18	1.63	1.12	1.26	0.91	0.80	0.94	0.97

Note. SE = standard error; CP = coverage rate; $\Delta =$ mode separation; D = number of documents; L = average document length.

Table 3.10

Standard Errors and Coverage Rates of $\boldsymbol{b}_2 = (b_{21}, b_{22}, b_{23}, b_{24})^T$ *Estimates Under the S1 Condition*

			SE				CP			
Δ	D	L	b_{21}	b_{22}	b_{23}	b_{24}	b_{21}	b_{22}	b_{23}	b_{24}
2.5	300	50	4.93	5.25	9.49	7.79	0.82	0.97	0.92	0.90
		100	3.46	3.77	4.35	4.55	0.92	0.97	0.94	0.93
		200	2.76	3.53	3.13	3.88	0.94	0.92	0.97	0.93
	500	50	3.33	3.41	4.34	4.91	0.91	0.91	0.94	0.78
		100	2.49	3.66	3.03	2.96	0.95	0.89	0.94	0.89
		200	2.44	2.87	2.62	2.84	0.88	0.98	0.94	0.91
	1000	50	2.18	2.46	2.58	2.89	0.85	0.72	0.93	0.75
		100	1.74	1.96	2.06	2.59	0.86	0.93	0.93	0.77
		200	1.50	1.81	1.87	1.83	0.92	0.96	0.92	0.91
3.5	300	50	4.13	4.08	7.25	7.24	0.91	0.96	0.94	0.87
		100	2.55	3.44	3.39	3.66	0.93	0.93	0.93	0.89
		200	1.95	2.31	2.07	2.53	0.95	0.97	0.97	0.94
	500	50	2.72	2.46	5.88	5.27	0.92	0.89	0.90	0.69
		100	1.97	2.21	2.23	2.73	0.88	0.90	0.93	0.86
		200	1.44	1.79	1.76	2.09	0.93	0.93	0.95	0.89
	1000	50	1.70	1.74	1.94	2.52	0.83	0.64	0.96	0.67
		100	1.24	1.48	1.46	1.85	0.84	0.81	0.93	0.71
		200	1.14	1.10	1.19	1.34	0.91	0.96	0.96	0.88

Note. SE = standard error; CP = coverage rate; $\Delta =$ mode separation; D = number of documents; L = average document length.

Table 3.11

Mean Absolute Bias, Relative Bias, Standard Errors, and Coverage Rates of σ^2 Estimates and Latent Class Membership Recovery Percentages Under the S1 Condition

Δ	D	L	AB	RB	SE	CP	Mem
2.5	300	50	10.60	-32.84	5.87	0.74	77.24
		100	4.67	-6.01	5.43	0.97	80.72
		200	3.40	-2.54	4.35	0.96	82.99
	500	50	8.47	-27.33	4.69	0.70	79.19
		100	4.54	-7.63	4.77	0.95	81.86
		200	2.77	1.22	3.60	0.96	83.24
	1000	50	6.32	-20.36	4.01	0.68	80.31
		100	2.78	-5.67	3.18	0.92	82.46
		200	2.36	-1.90	2.81	0.94	83.86
3.5	300	50	5.99	-39.75	1.91	0.74	80.77
		100	3.38	-17.14	3.01	0.91	84.76
		200	1.93	-7.57	2.09	0.96	88.03
	500	50	5.89	-38.81	2.33	0.54	81.55
		100	2.94	-15.87	2.38	0.84	85.77
		200	1.50	-3.71	1.77	0.95	88.10
	1000	50	5.21	-34.61	1.90	0.38	83.29
		100	2.32	-13.67	1.81	0.78	86.53
		200	1.32	-5.59	1.36	0.93	88.64

Note. $\Delta =$ mode separation; D = number of documents; L = average document length; AB = mean absolute bias; RB = mean relative bias; SE = standard error; CP = coverage rate; Mem = latent class membership recovery percentage

Table 3.12

Mean Absolute Bias and Relative Bias of $\boldsymbol{b}_1 = (b_{11}, b_{12}, b_{13}, b_{14})^T$ *Estimates Under the S2 Condition*

			AB				RB			
Δ	D	L	b_{11}	b_{12}	b_{13}	b_{14}	b_{11}	b_{12}	b_{13}	b_{14}
2.5	300	50	2.60	4.41	2.86	3.37	-2.33	4.29	-5.04	5.29
		100	2.23	2.75	2.25	2.57	4.51	-0.79	5.58	5.93
		200	2.00	3.13	2.35	2.74	0.10	-2.82	10.54	9.34
	500	50	1.91	3.18	2.07	2.62	-4.81	2.43	1.06	4.39
		100	1.59	2.15	1.77	2.33	-2.79	-0.95	5.21	8.14
		200	1.50	2.39	1.77	2.35	1.07	-2.26	7.07	7.80
	1000	50	1.37	2.32	1.13	1.40	-1.97	1.53	0.06	0.92
		100	1.12	1.60	1.43	1.65	-0.99	-0.24	4.18	3.93
		200	1.17	1.57	1.37	1.45	1.35	-1.51	5.83	2.91
3.5	300	50	1.91	3.52	1.73	2.10	-3.91	3.59	-0.39	3.92
		100	1.65	2.66	1.54	2.02	3.86	-2.98	6.85	6.07
		200	1.48	3.08	1.71	1.80	3.66	-4.19	10.14	6.34
	500	50	1.29	2.19	1.56	1.57	-0.71	1.78	2.26	0.88
		100	1.01	1.78	1.27	1.49	0.82	-1.54	4.28	5.10
		200	0.97	2.22	1.44	1.41	1.67	-2.85	7.46	4.65
	1000	50	0.99	1.73	0.95	1.13	-2.15	1.37	0.80	1.48
		100	0.77	1.27	1.05	1.09	-0.37	-0.58	4.74	2.63
		200	0.71	1.52	1.05	1.05	1.08	-1.83	5.56	3.21

Note. AB = mean absolute bias; RB = mean relative bias; $\Delta =$ mode separation; D = number of documents; L = average document length.

Table 3.13

Mean Absolute Bias and Relative Bias of $\boldsymbol{b}_2 = (b_{21}, b_{22}, b_{23}, b_{24})^T$ *Estimates Under the S2 Condition*

			AB				RB			
Δ	D	L	b_{21}	b_{22}	b_{23}	b_{24}	b_{21}	b_{22}	b_{23}	b_{24}
2.5	300	50	2.81	3.42	2.73	3.30	2.41	-2.47	2.09	-10.95
		100	2.07	3.47	2.21	2.65	0.38	-4.07	2.21	4.39
		200	1.83	3.30	2.07	2.61	-0.39	-4.06	2.03	4.43
	500	50	1.82	2.23	2.24	2.77	0.97	-0.91	0.04	-8.90
		100	1.64	2.08	1.73	2.05	-0.09	-2.21	2.00	-1.06
		200	1.65	2.26	1.76	1.81	-1.02	-2.43	3.04	2.29
	1000	50	1.39	1.65	1.45	2.06	1.10	0.52	0.90	-7.30
		100	1.07	1.52	1.40	1.27	0.48	-1.16	1.63	-2.05
		200	0.99	1.55	1.28	1.29	-0.45	-1.88	2.16	3.43
3.5	300	50	1.98	2.58	1.76	2.64	1.79	-1.91	1.59	-7.26
		100	1.58	2.23	1.61	2.28	-0.74	-2.31	1.94	6.01
		200	1.32	2.71	1.68	2.11	-1.06	-3.47	3.44	7.38
	500	50	1.30	1.48	1.90	2.09	0.52	-0.53	0.90	-5.49
		100	1.08	1.54	1.31	1.47	-0.46	-1.88	2.02	4.27
		200	1.25	1.85	1.18	1.50	-1.01	-2.46	2.25	6.18
	1000	50	0.96	1.16	0.95	1.74	0.63	0.63	0.98	-5.08
		100	0.79	1.01	1.14	0.93	-0.24	-0.61	1.84	-0.17
		200	0.72	1.32	0.95	1.11	-0.47	-1.53	2.28	3.82

Note. AB = mean absolute bias; RB = mean relative bias; $\Delta =$ mode separation; D = number of documents; L = average document length.

Table 3.14

Standard Errors and Coverage Rates of $\boldsymbol{b}_1 = (b_{11}, b_{12}, b_{13}, b_{14})^T$ *Estimates Under the S2 Condition*

			SE				CP			
Δ	D	L	b_{11}	b_{12}	b_{13}	b_{14}	b_{11}	b_{12}	b_{13}	b_{14}
2.5	300	50	3.19	4.50	3.52	4.14	0.96	0.93	0.92	0.93
		100	2.66	3.58	2.76	3.17	0.93	0.97	0.93	0.94
		200	2.53	3.28	2.43	3.03	0.92	0.88	0.91	0.89
	500	50	2.40	3.56	3.84	3.28	0.96	0.94	0.94	0.92
		100	1.93	2.70	2.14	2.47	0.94	0.94	0.93	0.90
		200	1.91	2.49	1.94	2.58	0.94	0.87	0.90	0.87
	1000	50	1.66	2.54	1.43	1.70	0.95	0.91	0.97	0.97
		100	1.52	1.93	1.63	1.90	0.91	0.96	0.92	0.93
		200	1.39	1.60	1.51	1.78	0.92	0.90	0.89	0.90
3.5	300	50	2.14	3.46	2.12	2.49	0.97	0.94	0.95	0.95
		100	1.94	2.62	1.68	2.58	0.93	0.84	0.94	0.92
		200	1.64	2.11	1.44	1.94	0.93	0.75	0.88	0.91
	500	50	1.77	2.39	3.44	2.03	0.95	0.96	0.96	0.96
		100	1.26	1.98	1.45	1.59	0.98	0.94	0.94	0.92
		200	1.27	1.74	1.28	1.54	0.92	0.76	0.82	0.90
	1000	50	1.14	1.90	1.16	1.35	0.96	0.91	0.95	0.95
		100	0.94	1.47	1.09	1.30	0.96	0.95	0.90	0.92
		200	0.88	1.32	1.02	1.14	0.94	0.80	0.83	0.89

Note. SE = standard error; CP = coverage rate; $\Delta =$ mode separation; D = number of documents; L = average document length.

Table 3.15

Standard Errors and Coverage Rates of $\boldsymbol{b}_2 = (b_{21}, b_{22}, b_{23}, b_{24})^T$ *Estimates Under the S2 Condition*

			SE				CP			
Δ	D	L	b_{21}	b_{22}	b_{23}	b_{24}	b_{21}	b_{22}	b_{23}	b_{24}
2.5	300	50	3.23	3.94	3.19	3.72	0.94	0.93	0.97	0.98
		100	2.65	3.33	2.78	3.19	0.98	0.90	0.89	0.96
		200	2.33	2.98	2.47	3.11	0.95	0.85	0.92	0.96
	500	50	2.24	2.91	3.63	3.29	0.93	0.96	0.95	0.95
		100	2.04	2.05	2.14	2.66	0.94	0.96	0.90	0.98
		200	1.97	2.31	1.89	2.27	0.90	0.89	0.93	0.96
	1000	50	1.59	2.01	1.74	1.92	0.95	0.94	0.96	0.97
		100	1.37	1.86	1.70	1.61	0.94	0.92	0.89	0.96
		200	1.23	1.29	1.42	1.47	0.94	0.90	0.90	0.95
3.5	300	50	2.45	3.01	2.11	2.94	0.91	0.94	0.96	0.96
		100	1.96	2.27	1.91	2.61	0.95	0.90	0.93	0.91
		200	1.57	2.21	1.70	2.16	0.93	0.77	0.89	0.86
	500	50	1.64	1.77	3.55	2.44	0.93	0.97	0.91	0.94
		100	1.40	1.56	1.48	1.62	0.93	0.92	0.93	0.97
		200	1.42	1.39	1.21	1.42	0.88	0.84	0.91	0.88
	1000	50	1.14	1.41	1.14	1.86	0.95	0.93	0.97	0.86
		100	1.05	1.25	1.23	1.22	0.93	0.93	0.87	0.94
		200	0.83	1.13	0.86	1.12	0.98	0.79	0.86	0.92

Note. SE = standard error; CP = coverage rate; $\Delta =$ mode separation; D = number of documents; L = average document length.

Table 3.16

Mean Absolute Bias, Relative Bias, Standard Errors, and Coverage Rates of σ^2 Estimates and Latent Class Membership Recovery Percentage Under the S2 Condition

Δ	D	L	AB	RB	SE	CP	Mem
2.5	300	50	6.14	-17.65	4.93	0.80	81.16
		100	4.08	0.13	5.20	0.90	82.58
		200	2.95	2.09	3.71	0.97	84.19
	500	50	4.55	-11.76	4.05	0.86	81.76
		100	2.89	0.41	3.59	0.98	83.74
		200	2.62	1.62	3.18	0.95	84.70
	1000	50	3.21	-9.67	2.91	0.80	82.98
		100	1.94	0.35	2.50	0.95	84.22
		200	1.72	0.07	2.11	0.98	85.18
3.5	300	50	4.86	-30.70	2.61	0.69	85.84
		100	2.39	-7.52	2.68	0.89	88.44
		200	1.62	-0.28	2.01	0.95	89.89
	500	50	3.83	-23.46	2.59	0.61	86.09
		100	1.47	-5.09	1.69	0.94	88.83
		200	1.37	-1.93	1.62	0.96	90.05
	1000	50	2.72	-18.07	1.46	0.61	87.03
		100	1.18	-5.17	1.19	0.94	89.01
		200	0.89	-1.34	1.14	0.93	90.40

Note. $\Delta =$ mode separation; D = number of documents; L = average document length; AB = mean absolute bias; RB = mean relative bias; SE = standard error; CP = coverage rate; Mem = latent class membership recovery percentage

Table 3.17

Result Summary for $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ under the S1 Condition

			$\boldsymbol{b}_1$				$\boldsymbol{b}_2$			
			Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
Δ	D	L	AB	RB	SE	CP	AB	RB	SE	CP
2.5	300	50	6.08	9.30	6.06	0.93	6.80	15.30	6.87	0.90
		100	3.98	8.27	3.94	0.94	3.62	6.02	4.03	0.94
		200	2.89	3.62	3.38	0.92	2.77	3.12	3.33	0.94
	500	50	4.75	10.54	4.21	0.85	4.29	10.55	4.00	0.89
		100	2.99	5.80	2.97	0.89	2.83	5.26	3.03	0.92
		200	2.41	3.59	2.65	0.92	2.25	2.76	2.69	0.93
	1000	50	3.00	7.85	2.53	0.86	3.21	8.41	2.53	0.81
		100	2.18	5.40	1.94	0.87	2.10	4.44	2.09	0.87
		200	1.79	3.03	1.95	0.91	1.55	2.44	1.75	0.93
3.5	300	50	4.52	8.85	4.44	0.93	4.99	8.96	5.67	0.92
		100	3.05	4.93	3.26	0.92	3.01	5.35	3.26	0.92
		200	2.08	2.38	2.56	0.93	1.85	2.03	2.21	0.96
	500	50	3.81	6.71	4.37	0.85	3.96	9.67	4.08	0.85
		100	2.29	4.38	2.23	0.90	2.14	4.06	2.28	0.89
		200	1.68	2.91	1.81	0.93	1.46	1.86	1.77	0.93
	1000	50	2.78	6.76	2.16	0.81	2.67	7.09	1.98	0.78
		100	1.71	4.07	1.53	0.89	1.74	4.31	1.51	0.82
		200	1.23	2.14	1.30	0.91	1.07	1.71	1.19	0.93

Note. Mean AB = averaged mean absolute bias across regression coefficients; Mean RB = averaged mean relative bias across regression coefficients; Mean SE = averaged standard error across regression coefficients; Mean CP = averaged coverage rate across regression coefficients; $\Delta =$ mode separation; D = number of documents; L = average document length.

Table 3.18

Result Summary for $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ under the S2 Condition

			$\boldsymbol{b}_1$				$\boldsymbol{b}_2$			
			Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
Δ	D	L	AB	RB	SE	CP	AB	RB	SE	CP
2.5	300	50	3.31	4.24	3.84	0.94	3.06	4.48	3.52	0.96
		100	2.45	4.20	3.04	0.94	2.60	2.76	2.99	0.93
		200	2.56	5.70	2.82	0.90	2.45	2.73	2.72	0.92
	500	50	2.45	3.18	3.27	0.94	2.26	2.70	3.02	0.95
		100	1.96	4.27	2.31	0.93	1.88	1.34	2.22	0.95
		200	2.00	4.55	2.23	0.90	1.87	2.20	2.11	0.92
	1000	50	1.56	1.12	1.83	0.95	1.64	2.46	1.82	0.96
		100	1.45	2.33	1.75	0.93	1.31	1.33	1.63	0.93
		200	1.39	2.90	1.57	0.90	1.27	1.98	1.35	0.92
3.5	300	50	2.32	2.95	2.55	0.95	2.24	3.14	2.63	0.94
		100	1.97	4.94	2.21	0.91	1.93	2.75	2.19	0.92
		200	2.02	6.08	1.78	0.87	1.95	3.84	1.91	0.86
	500	50	1.65	1.41	2.41	0.96	1.69	1.86	2.35	0.94
		100	1.39	2.93	1.57	0.95	1.35	2.16	1.51	0.94
		200	1.51	4.16	1.46	0.85	1.44	2.97	1.36	0.88
	1000	50	1.20	1.45	1.39	0.94	1.20	1.83	1.39	0.93
		100	1.04	2.08	1.20	0.93	0.97	0.71	1.19	0.92
		200	1.08	2.92	1.09	0.87	1.02	2.03	0.98	0.89

Note. Mean AB = averaged mean absolute bias across regression coefficients; Mean RB = averaged mean relative bias across regression coefficients; Mean SE = averaged standard error across regression coefficients; Mean CP = averaged coverage rate across regression coefficients; $\Delta =$ mode separation; D = number of documents; L = average document length.

Table 4.1

The Number of Documents and the Average Document Length for Each Time Point

	Time 1	Time 2	Time 3	Time 4
Number of documents	115	136	111	131
Average document length	38.78 (20.57)	52.65 (19.79)	55.14 (19.00)	59.42 (20.97)

Note. The numbers in the parentheses are standard deviations.

Table 4.2

DIC and AICM Values for LDA Models with Two to Six Topics

K	DIC	AICM
2	279054	-279846
3	277103	**-278931**
4	276162	-279049
5	275485	-279607
6	**274992**	-280092

Table 4.3

The 15 Words Having the Highest Posterior Mean Estimates of Probabilities.

	Topic 1		Topic 2		Topic 3	
	Word	γ_{kv}	Word	γ_{kv}	Word	γ_{kv}
1	put	0.028	energy	0.047	change	0.065
2	stronger	0.025	increase	0.025	variable	0.063
3	big	0.019	population	0.024	independent	0.029
4	think	0.017	decrease	0.024	cause	0.027
5	little	0.017	different	0.021	dependent	0.020
6	bigger	0.016	amount	0.021	effect	0.019
7	hotter	0.014	kinetic	0.019	different	0.017
8	all	0.014	temperature	0.018	think	0.016
9	bottlea	0.013	time	0.018	experiment	0.016
10	lot	0.013	person	0.018	hypothesis	0.015
11	muscle	0.012	potential	0.018	bottlea	0.012
12	start	0.012	disease	0.018	fly	0.011
13	hot	0.011	pot	0.017	hour	0.011
14	people	0.010	hypothesis	0.016	happen	0.010
15	down	0.010	hotter	0.015	faster	0.010

Table 4.4

Average of Posterior Means of Proportions for Each of the Topics

	Topic 1	Topic 2	Topic 3
Time 1 (N=115)	0.47	0.25	0.28
Time 2 (N=136)	0.38	0.36	0.25
Time 3 (N=111)	0.36	0.41	0.23
Time 4 (N=131)	0.30	0.50	0.20

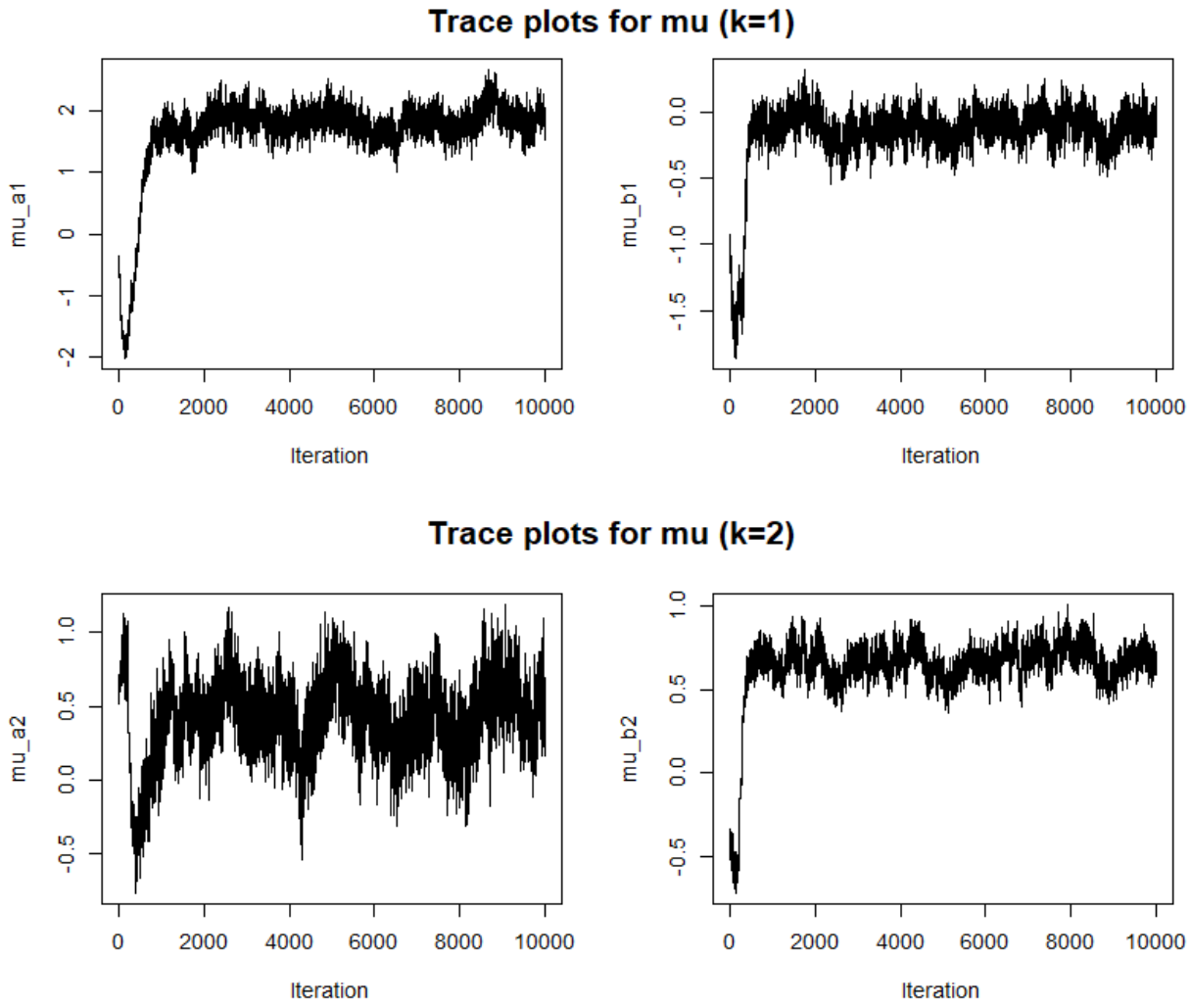

Figure 4.1. Trace plots for $\boldsymbol{\mu}_{\xi_1}$ and $\boldsymbol{\mu}_{\xi_2}$.

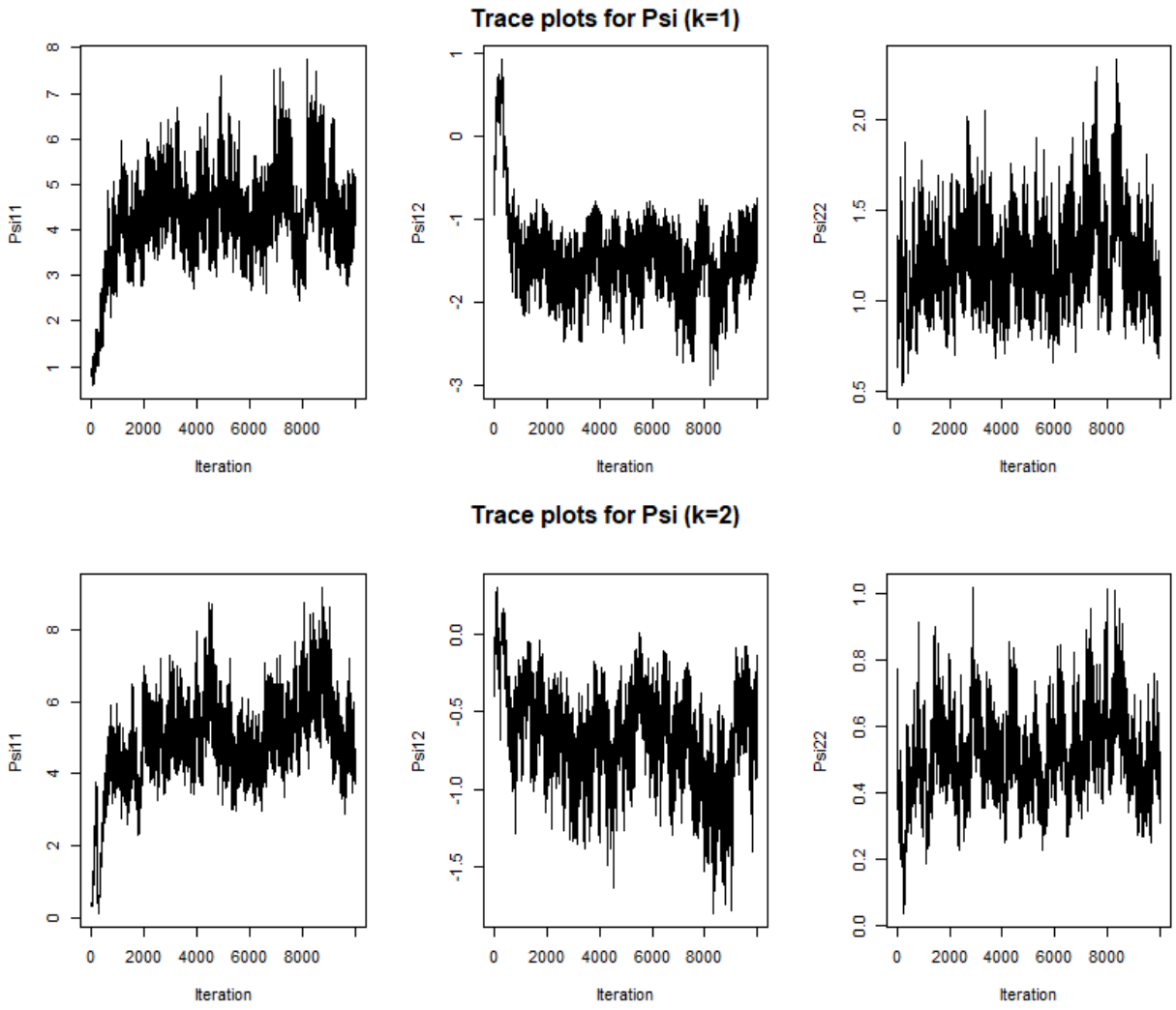

Figure 4.2. Trace plots for $\mathbf{\Psi}_1$ and $\mathbf{\Psi}_2$.

Table 4.5

The 15 Words Having the Highest Posterior Mean Estimates of Probabilities

	Topic 1		Topic 2		Topic 3	
	Word	γ_{kv}	Word	γ_{kv}	Word	γ_{kv}
1	put	0.024	energy	0.043	variable	0.103
2	stronger	0.023	increase	0.023	change	0.068
3	think	0.020	population	0.022	independent	0.048
4	big	0.016	decrease	0.022	dependent	0.033
5	change	0.016	different	0.020	experiment	0.018
6	bigger	0.016	amount	0.019	hypothesis	0.018
7	little	0.014	kinetic	0.017	cause	0.016
8	bottleA	0.014	hypothesis	0.017	effect	0.015
9	all	0.013	disease	0.016	response	0.015
10	hotter	0.012	temperature	0.016	manipulate	0.012
11	cause	0.012	potential	0.016	different	0.012
12	lot	0.012	time	0.016	know	0.011
13	fly	0.011	pot	0.016	happen	0.010
14	hot	0.010	person	0.015	observation	0.010
15	muscle	0.010	hotter	0.014	celsius	0.009

Table 4.6

Posterior Mean Estimates of the Parameters in the Structural Model Component

	Parameter	Estimate
Topic 1	$\mu_{\xi_{a1}}$	1.84 (1.43, 2.26)
	$\mu_{\xi_{b1}}$	-0.12 (-0.33, 0.08)
	$\boldsymbol{\Psi}_1(1,1)$	4.42 (3.16, 6.15)
	$\boldsymbol{\Psi}_1(2,2)$	1.22 (0.86, 1.76)
	$\boldsymbol{\Psi}_1(1,2)$	-1.52 (-2.26, -0.98)
Topic 2	$\mu_{\xi_{a2}}$	0.44 (-0.01, 0.87)
	$\mu_{\xi_{b2}}$	0.68 (0.50, 0.86)
	$\boldsymbol{\Psi}_2(1,1)$	5.20 (3.65, 7.31)
	$\boldsymbol{\Psi}_2(2,2)$	0.51 (0.31, 0.76)
	$\boldsymbol{\Psi}_2(1,2)$	-0.71 (-1.31, -0.23)

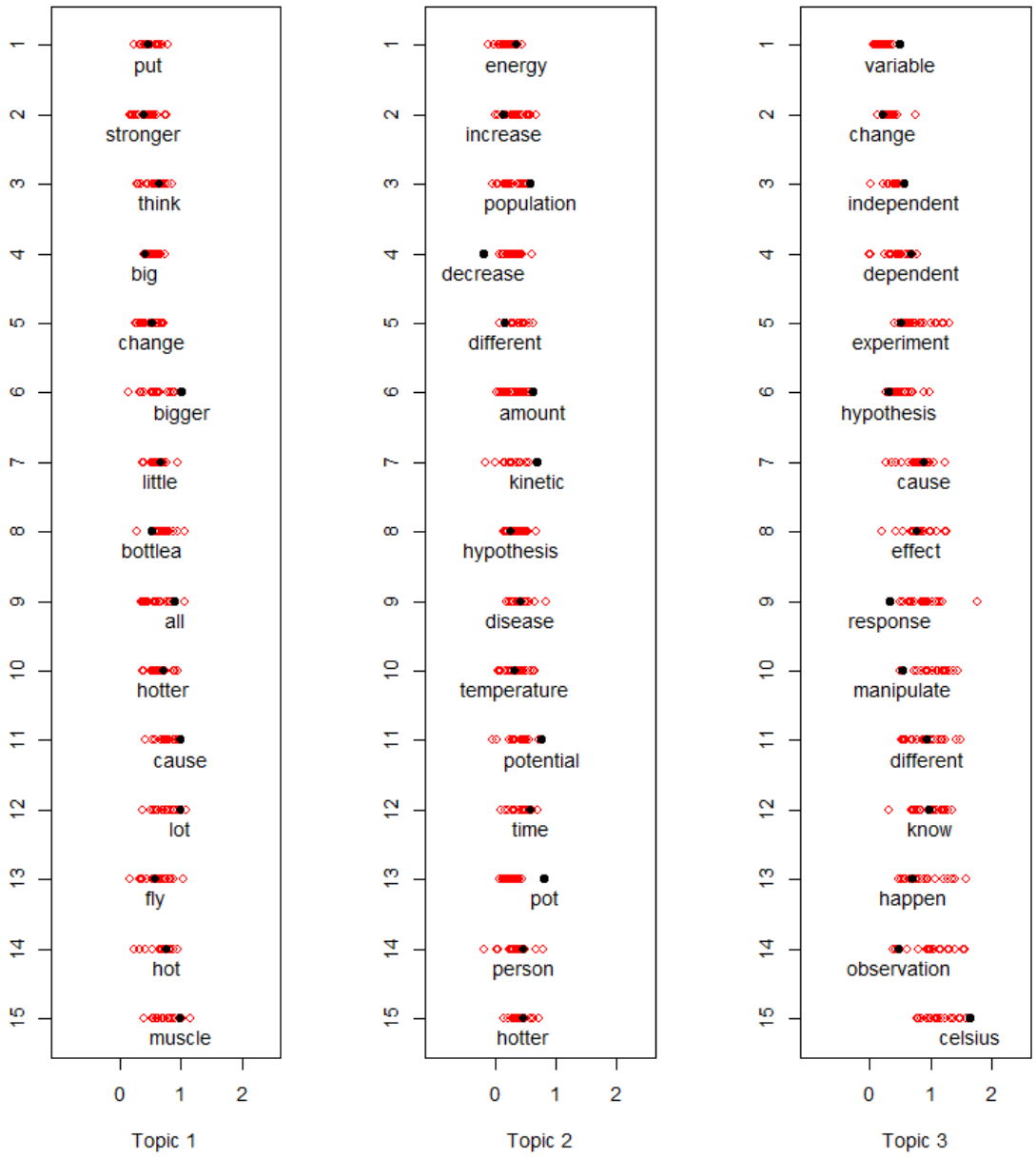

Figure 4.3. IMI scores for words in each topic.

Table 4.7

Summary of Simulation Study Conditions

	D	L	$\mu_{\xi_{a1}}$	$\mu_{\xi_{b1}}$	$\mu_{\xi_{a2}}$	$\mu_{\xi_{b2}}$	$\boldsymbol{\Psi}_k(1,1)$	$\boldsymbol{\Psi}_k(2,2)$	$\boldsymbol{\Psi}_k(1,2)$
V1	150	50	1.5	-0.2	0.4	0.5	3	0.5	-0.73
		100	1.5	-0.2	0.4	0.5	3	0.5	-0.73
		150	1.5	-0.2	0.4	0.5	3	0.5	-0.73
	250	50	1.5	-0.2	0.4	0.5	3	0.5	-0.73
		100	1.5	-0.2	0.4	0.5	3	0.5	-0.73
		150	1.5	-0.2	0.4	0.5	3	0.5	-0.73
	500	50	1.5	-0.2	0.4	0.5	3	0.5	-0.73
		100	1.5	-0.2	0.4	0.5	3	0.5	-0.73
		150	1.5	-0.2	0.4	0.5	3	0.5	-0.73
V2	150	50	1.5	-0.2	0.4	0.5	5	1	-1.34
		100	1.5	-0.2	0.4	0.5	5	1	-1.34
		150	1.5	-0.2	0.4	0.5	5	1	-1.34
	250	50	1.5	-0.2	0.4	0.5	5	1	-1.34
		100	1.5	-0.2	0.4	0.5	5	1	-1.34
		150	1.5	-0.2	0.4	0.5	5	1	-1.34
	500	50	1.5	-0.2	0.4	0.5	5	1	-1.34
		100	1.5	-0.2	0.4	0.5	5	1	-1.34
		150	1.5	-0.2	0.4	0.5	5	1	-1.34

Note. D = number of documents for each time point. L = average document length.

Table 4.8

Mean Absolute Bias, Standard Errors, and Coverage Rates for $\boldsymbol{\mu}_{\xi_1} = \left(\mu_{\xi_{a1}}, \mu_{\xi_{b1}}\right)^T$

			$\mu_{\xi_{a1}}$			$\mu_{\xi_{b1}}$		
Var	D	L	AB	SE	CvR	AB	SE	CvR
V1	150	50	0.16	0.19	0.96	0.08	0.08	0.87
		100	0.12	0.14	0.96	0.06	0.07	0.90
		150	0.11	0.14	0.98	0.07	0.07	0.89
	250	50	0.12	0.15	0.93	0.06	0.06	0.89
		100	0.10	0.12	0.94	0.05	0.05	0.91
		150	0.10	0.12	0.95	0.04	0.05	0.94
	500	50	0.07	0.09	0.97	0.04	0.04	0.94
		100	0.07	0.09	0.91	0.04	0.04	0.91
		150	0.08	0.08	0.90	0.03	0.04	0.90
V2	150	50	0.16	0.22	0.92	0.09	0.10	0.93
		100	0.15	0.19	0.95	0.09	0.11	0.89
		150	0.15	0.19	0.94	0.07	0.09	0.94
	250	50	0.12	0.15	0.97	0.07	0.08	0.91
		100	0.12	0.14	0.95	0.06	0.06	0.95
		150	0.12	0.14	0.94	0.05	0.06	0.93
	500	50	0.09	0.10	0.96	0.05	0.05	0.93
		100	0.10	0.10	0.90	0.05	0.06	0.89
		150	0.11	0.11	0.91	0.04	0.05	0.92

Note. Var = variance condition. D = number of documents for each time point. L = average document length. AB = mean absolute bias. SE = standard error. CvR = coverage rate.

Table 4.9

Mean Absolute Bias, Standard Errors, and Coverage Rates for $\boldsymbol{\mu}_{\xi_2} = \left(\mu_{\xi_{a2}}, \mu_{\xi_{b2}}\right)^T$

			$\mu_{\xi_{a2}}$			$\mu_{\xi_{b2}}$		
Var	D	L	AB	SE	CvR	AB	SE	CvR
V1	150	50	0.24	0.22	0.83	0.06	0.08	0.96
		100	0.13	0.15	0.96	0.04	0.06	0.99
		150	0.13	0.17	0.92	0.06	0.07	0.90
	250	50	0.13	0.16	0.95	0.05	0.06	0.96
		100	0.11	0.14	0.94	0.04	0.06	0.94
		150	0.09	0.11	0.99	0.05	0.05	0.93
	500	50	0.09	0.11	0.95	0.03	0.04	0.98
		100	0.07	0.08	0.98	0.03	0.04	0.89
		150	0.07	0.08	0.97	0.03	0.03	0.92
V2	150	50	0.18	0.22	0.94	0.08	0.10	0.95
		100	0.14	0.18	0.95	0.07	0.08	0.96
		150	0.17	0.21	0.93	0.07	0.09	0.96
	250	50	0.14	0.18	0.95	0.07	0.08	0.94
		100	0.13	0.16	0.97	0.06	0.07	0.96
		150	0.14	0.17	0.92	0.06	0.08	0.92
	500	50	0.08	0.10	0.98	0.05	0.05	0.94
		100	0.08	0.10	0.96	0.05	0.05	0.93
		150	0.08	0.10	0.96	0.04	0.05	0.93

Note. Var = variance condition. D = number of documents for each time point. L = average document length. AB = mean absolute bias. SE = standard error. CvR = coverage rate.

Table 4.10

Mean Absolute Bias, Standard Errors, and Coverage Rates for $\boldsymbol{\Psi}_1$

			$\boldsymbol{\Psi}_1(1,1)$			$\boldsymbol{\Psi}_1(1,2)$			$\boldsymbol{\Psi}_1(2,2)$		
Var	D	L	AB	SE	CvR	AB	SE	CvR	AB	SE	CvR
V1	150	50	0.45	0.57	0.99	0.18	0.21	0.97	0.08	0.11	0.96
		100	0.40	0.38	0.91	0.13	0.13	0.91	0.07	0.07	0.94
		150	0.47	0.40	0.84	0.15	0.13	0.90	0.07	0.06	0.90
	250	50	0.36	0.41	0.93	0.13	0.15	0.94	0.07	0.07	0.93
		100	0.39	0.36	0.86	0.13	0.13	0.84	0.07	0.06	0.78
		150	0.37	0.30	0.86	0.11	0.11	0.86	0.06	0.05	0.84
	500	50	0.36	0.28	0.77	0.10	0.10	0.87	0.06	0.05	0.86
		100	0.37	0.24	0.63	0.10	0.08	0.81	0.06	0.04	0.78
		150	0.36	0.21	0.64	0.09	0.08	0.78	0.06	0.04	0.78
V2	150	50	0.76	0.80	0.98	0.29	0.32	0.98	0.15	0.17	0.91
		100	0.77	0.67	0.85	0.25	0.25	0.90	0.15	0.15	0.87
		150	0.78	0.63	0.86	0.26	0.26	0.86	0.15	0.13	0.84
	250	50	0.62	0.66	0.88	0.22	0.24	0.91	0.13	0.12	0.91
		100	0.76	0.58	0.74	0.23	0.21	0.85	0.14	0.11	0.81
		150	0.68	0.46	0.82	0.17	0.17	0.90	0.11	0.09	0.91
	500	50	0.72	0.52	0.70	0.21	0.19	0.78	0.14	0.11	0.74
		100	0.72	0.34	0.66	0.19	0.12	0.76	0.13	0.07	0.64
		150	0.66	0.35	0.61	0.18	0.14	0.80	0.12	0.08	0.69

Note. Var = variance condition. D = number of documents for each time point. L = average document length. AB = mean absolute bias. SE = standard error. CvR = coverage rate.

Table 4.11

Mean Absolute Bias, Standard Errors, and Coverage Rates for $\boldsymbol{\Psi}_2$

			$\boldsymbol{\Psi}_2(1,1)$			$\boldsymbol{\Psi}_2(1,2)$			$\boldsymbol{\Psi}_2(2,2)$		
Var	D	L	AB	SE	CvR	AB	SE	CvR	AB	SE	CvR
V1	150	50	0.43	0.58	0.95	0.16	0.20	0.92	0.08	0.10	0.96
		100	0.42	0.48	0.97	0.13	0.15	0.92	0.06	0.07	0.97
		150	0.35	0.42	0.91	0.13	0.15	0.92	0.05	0.06	0.98
	250	50	0.41	0.49	0.92	0.13	0.16	0.92	0.06	0.07	0.96
		100	0.30	0.36	0.92	0.10	0.11	0.94	0.05	0.06	0.93
		150	0.34	0.33	0.85	0.10	0.11	0.88	0.05	0.06	0.91
	500	50	0.27	0.31	0.89	0.09	0.10	0.92	0.04	0.05	0.96
		100	0.24	0.23	0.93	0.07	0.08	0.93	0.04	0.04	0.96
		150	0.30	0.28	0.83	0.09	0.10	0.77	0.05	0.04	0.83
V2	150	50	0.71	0.87	0.95	0.32	0.40	0.89	0.17	0.22	0.91
		100	0.71	0.81	0.92	0.25	0.29	0.92	0.12	0.15	0.95
		150	0.59	0.64	0.91	0.21	0.24	0.94	0.12	0.13	0.94
	250	50	0.62	0.71	0.94	0.24	0.26	0.93	0.12	0.14	0.90
		100	0.53	0.57	0.87	0.18	0.20	0.87	0.12	0.12	0.89
		150	0.66	0.59	0.78	0.20	0.20	0.86	0.11	0.11	0.88
	500	50	0.53	0.51	0.85	0.16	0.16	0.88	0.09	0.08	0.92
		100	0.47	0.38	0.85	0.14	0.13	0.95	0.09	0.08	0.87
		150	0.51	0.42	0.78	0.15	0.15	0.81	0.09	0.08	0.80

Note. Var = variance condition. D = number of documents for each time point. L = average document length. AB = mean absolute bias. SE = standard error. CvR = coverage rate.

APPENDIX A

CALCULATIONS FOR EQUATION 2.10

Appendix A describes the computation of covariance between observed scores from two parallel tests (X_j and $\tilde{X}_{j'}$), which is represented as $Cov(X_j, \tilde{X}_{j'})$. This covariance can be expressed by using the expectations of X_j and $\tilde{X}_{j'}$:

$$Cov(X_j, \tilde{X}_{j'}) = E(X_j \tilde{X}_{j'}) - E(X_j)E(\tilde{X}_{j'}). \qquad (B1)$$

Suppose item j has C_j categories, and item j' has $C_{j'}$ categories. The vectors of underlying continuous variables $\boldsymbol{X}^* = (X_1^*, \dots, X_J^*)$ and $\tilde{\boldsymbol{X}}^* = (\tilde{X}_1^*, \dots, \tilde{X}_J^*)$ are assumed to follow a multivariate normal distribution with covariance matrix Σ that has 1 as the diagonal elements. The first term on the right hand side of Equation (B1) can be expressed as

$$E(X_j \tilde{X}_{j'}) = \sum_{k=0}^{C_j - 1} \sum_{l=0}^{C_{j'} - 1} klP\left(v_{j_k} < X_j^* \le v_{j_{k+1}},\ h_{j'_l} < \tilde{X}_{j'}^* \le h_{j'_{l+1}}\right), \quad (B2)$$

where $\{v_{j_0}, v_{j_1}, \dots, v_{j_{C_j}}\}$ and $\{h_{j'_0}, h_{j'_1}, \dots, h_{j'_{C_{j'}}}\}$ are thresholds for items j and j', respectively, and v_{j_0}, $h_{j'_0}$, $v_{j_{C_j}}$, and $h_{j'_{C_{j'}}}$ are $-\infty$, $-\infty$, ∞, and ∞, respectively. Since $X_{j'}^*$ and $\tilde{X}_{j'}^*$ are the underlying continuous variables from the parallel items, Equation (B2) can be rewritten as

$$\begin{aligned} E(X_j \tilde{X}_{j'}) &= \sum_{k=0}^{C_j - 1} \sum_{l=0}^{C_{j'} - 1} klP\left(v_{j_k} < X_j^* \le v_{j_{k+1}},\ h_{j'_l} < X_{j'}^* \le h_{j'_{l+1}}\right) \\ &= \sum_{k=0}^{C_j - 1} \sum_{l=0}^{C_{j'} - 1} kl\left[\Phi_2\left(v_{j_{k+1}}, h_{j'_{l+1}};\ \rho_{X_j^* X_{j'}^*}\right) - \Phi_2\left(v_{j_{k+1}}, h_{j'_l}; \rho_{X_j^* X_{j'}^*}\right)\right. \\ &\quad \left. - \Phi_2\left(v_{j_k}, h_{j'_{l+1}};\ \rho_{X_j^* X_{j'}^*}\right) + \Phi_2\left(v_{j_k}, h_{j'_l}; \rho_{X_j^* X_{j'}^*}\right)\right] \end{aligned}$$

$$= \left\{\sum_{k=1}^{C_j-1}\sum_{l=1}^{C_{j'}-1}(kl-k-l+1)\,\Phi_2\left(\nu_{j_k}, h_{j'_l};\ \rho_{X_j^* X_{j'}^*}\right) + \sum_{k=1}^{C_j-1}\big(kC_{j'} - C_{j'} - k + 1\big)\Phi_2\left(\nu_{j_k}, h_{j'_{C_{j'}}};\ \rho_{X_j^* X_{j'}^*}\right) + \sum_{l=1}^{C_{j'}-1}\big(C_j l - l - C_j + 1\big)\,\Phi_2\left(\nu_{j_{C_j}}, h_{j'_l};\ \rho_{X_j^* X_{j'}^*}\right) + \big(C_j C_{j'} - C_j - C_{j'} + 1\big)\Phi_2\left(\nu_{j_{C_j}}, h_{j'_{C_{j'}}};\ \rho_{X_j^* X_{j'}^*}\right)\right\} - \left\{\sum_{k=1}^{C_j-1}\sum_{l=1}^{C_{j'}-1}(kl-l)\,\Phi_2\left(\nu_{j_k}, h_{j'_l};\ \rho_{X_j^* X_{j'}^*}\right) + \sum_{l=1}^{C_{j'}-1}\big(C_j l - l\big)\Phi_2\left(\nu_{j_{C_j}}, h_{j'_l};\ \rho_{X_j^* X_{j'}^*}\right)\right\} -$$

$$\left\{\sum_{k=1}^{C_j-1}\sum_{l=1}^{C_{j'}-1}(kl-k)\,\Phi_2\left(\nu_{j_k}, h_{j'_l}; \rho_{X_j^* X_{j'}^*}\right) + \sum_{k=1}^{C_j-1}\big(kC_{j'} - k\big)\Phi_2\left(\nu_{j_k}, h_{j'_{C_{j'}}};\ \rho_{X_j^* X_{j'}^*}\right)\right\} + \left\{\sum_{k=1}^{C_j-1}\sum_{l=1}^{C_{j'}-1} kl\,\Phi_2\left(\nu_{j_k}, h_{j'_l};\ \rho_{X_j^* X_{j'}^*}\right)\right\}$$

$$= \sum_{k=1}^{C_j-1}\sum_{l=1}^{C_{j'}-1}\Phi_2\left(\nu_{j_k}, h_{j'_l};\ \rho_{X_j^* X_{j'}^*}\right) - \big(C_{j'}-1\big)\sum_{k=1}^{C_j-1}\Phi_1\big(\nu_{j_k}\big) - \big(C_j-1\big)\sum_{l=1}^{C_{j'}-1}\Phi_1\left(h_{j'_l}\right) + \big(C_j-1\big)\big(C_{j'}-1\big),$$

where $\Phi_2\left(\nu_{j_k}, h_{j'_l}; \rho_{X_j^* X_{j'}^*}\right)$ is the cumulative bivariate normal distribution function of ν_{j_k} and $h_{j'_l}$ with correlation $\rho_{X_j^* X_{j'}^*}$, and $\Phi_1(\nu_{j_k})$ is the cumulative univariate normal distribution function of ν_{j_k}. Similarly, the expectation of X_j can be represented as

$$E\big(X_j\big) = \sum_{k=0}^{C_j-1} k\,P\big(X_j = k\big) = \sum_{k=0}^{C_j-1} kP\big(\nu_{j_k} < X_j^* \le \nu_{j_{k+1}}\big) = \sum_{k=0}^{C_j-1} k\big[\Phi_1\big(\nu_{j_{k+1}}\big) - \Phi_1\big(\nu_{j_k}\big)\big]$$

$$= \sum_{k=1}^{C_j}(k-1)\Phi_1\big(\nu_{j_k}\big) - \sum_{k=0}^{C_j-1} k\Phi_1\big(\nu_{j_k}\big)$$

$$= \sum_{k=1}^{C_j-1}(k-1)\Phi_1\big(\nu_{j_k}\big) - \sum_{k=1}^{C_j-1} k\Phi_1\big(\nu_{j_k}\big) + \big(C_j-1\big)\Phi_1\left(\nu_{j_{C_j}}\right) = -\sum_{k=1}^{C_j-1}\Phi_1\big(\nu_{j_k}\big) + \big(C_j-1\big).$$

Thus,

$$E(X_j)E(\tilde{X}_{j'}) = \sum_{k=1}^{C_j-1} \Phi_1(v_{j_k}) \sum_{l=1}^{C_{j'}-1} \Phi_1\left(h_{j_l'}\right) - (C_j - 1) \sum_{l=1}^{C_{j'}-1} \Phi_1\left(h_{j_l'}\right) - (C_{j'} - 1) \sum_{k=1}^{C_j-1} \Phi_1(v_{j_k})$$

$$+ (C_j - 1)(C_{j'} - 1).$$

Therefore,

$$Cov(X_j, \tilde{X}_{j'}) = E(X_j\tilde{X}_{j'}) - E(X_j)E(\tilde{X}_{j'})$$

$$= \sum_{k=1}^{C_j-1} \sum_{l=1}^{C_{j'}-1} \Phi_2\left(v_{j_k}, h_{j_l'}; \rho_{X_j^* X^*_{j'}}\right) - \sum_{k=1}^{C_j-1} \Phi_1(v_{j_k}) \sum_{l=1}^{C_{j'}-1} \Phi_1\left(h_{j_l'}\right).$$

APPENDIX B

SAMPLE R CODE FOR ESTIMATING NONLINEAR SEM RELIABILITY COEFFICIENT

The following is a program in R used for estimating nonlinear SEM reliability coefficient using the data generated from Structure1 with condition C25.

```
library(pbivnorm)
nfact = 1 # Number of factors
nitem = 9 # Number of items
nth = c(1,1,4,1,1,4,1,1,4)  # Number of thresholds for each item
maxths = max(nth) # Maximum number of thresholds

# Factor loading estimates
lamb.est = matrix(c(0.690,0.716,0.702,0.700, 0.636, 0.651, 0.643, 0.743, 0.69
1), ncol=nfact)

# Thresholds for items
Thr = matrix(c(-0.095,rep(NA,3),-0.015,rep(NA,3),-1.645,-0.548,0.681,1.645,
                0.005,rep(NA,3),0.015,rep(NA,3),-1.626,-0.577,0.752,1.728,
                0.010,rep(NA,3),0.111,rep(NA,3),-1.626,-0.668,0.559,1.774),nc
ol=nitem)

# Estimated polychoric correlations
polych = diag(nitem)
polych[lower.tri(polych,diag=F)] = c(0.537,0.455,0.417,0.474,0.422,0.423,0.51
4,0.522,
0.544,0.493,0.471,0.478,0.454,0.458,0.457,
0.487,0.456,0.420,0.445,0.549,0.478,0.448,0.422,0.409,0.596,0.512,
0.446,0.350,0.432,0.420,0.497,0.469,0.440,0.461,0.445,0.514)
polych[upper.tri(polych, diag=F)] = t(polych)[upper.tri(t(polych), diag=F)]

# Factor covariance matrix
FacCor = diag(nfact)

rhoM = lamb.est%*%FacCor%*%t(lamb.est)

#Calculation of nonlinear SEM reliability
nom = denom = matrix(NA, nrow=nitem,ncol=nitem)
for(j in 1:nitem){
for(jp in 1:nitem){
    phi2_n = phi2_d = as.data.frame(matrix(NA, nrow = nth[j], ncol = nth[jp])
)
    for(k in 1:nth[j]){
    for(l in 1:nth[jp]){
        phi2_n[k,l] = pbivnorm(Thr[k,j],Thr[l,jp], rho=rhoM[j,jp])
        phi2_d[k,l] = pbivnorm(Thr[k,j],Thr[l,jp], rho=polych[j,jp])
    } #l
```

```
    } #k
    phi1.j = pnorm(Thr[!is.na(Thr[,j]),j],0,1)
    phi1.jp = pnorm(Thr[!is.na(Thr[,jp]),jp],0,1)
    nom[j,jp] = sum(phi2_n) - sum(phi1.j)*sum(phi1.jp)
    denom[j,jp] = sum(phi2_d) - sum(phi1.j)*sum(phi1.jp)
} #jp
} #j

#nonlinear SEM reliability estimate
sem.rel = sum(nom)/sum(denom)
```

APPENDIX C

MISCELLANEOUS EQUATIONS

C.1 Latent Dirichlet Allocation

The generative model of LDA is as follows:

1. For each topic k, draw a $V \times 1$ vector of word distribution $\boldsymbol{\gamma}_k \sim Dirichlet(\boldsymbol{\beta})$.
2. For each document i,
    a. Draw a $K \times 1$ vector of topic proportions $\boldsymbol{\theta}_i \sim Dirichlet(\boldsymbol{\alpha})$.
3. For each word in document i ($w_{i,n}$),
    a. Assign a topic $z_{i,n} \sim Multinomial(\boldsymbol{\theta}_i)$.
    b. Draw a term $w_{i,n} \sim Multinomial(\boldsymbol{\gamma}_{z_{i,n}})$.

With augmented data $\boldsymbol{w}$ with realized values of $\boldsymbol{z}$, the joint distribution of $\boldsymbol{w}$, $\boldsymbol{\theta}$, $\boldsymbol{z}$, $\boldsymbol{\gamma}$ given the hyper parameters is

$$p(\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{\gamma}, |\boldsymbol{\alpha}, \boldsymbol{\beta}) = \left[\prod_{i=1}^{D} p(\boldsymbol{\theta}_i|\boldsymbol{\alpha}) \left\{\prod_{n=1}^{N_i} p(z_{i,n}|\boldsymbol{\theta}_i) p(w_{i,n}|z_{i,n}, \boldsymbol{\gamma})\right\}\right]\left[\prod_{k=1}^{K} p(\boldsymbol{\gamma}_k|\boldsymbol{\beta})\right]. \quad (C.1)$$

C.2 Gibbs sampling equations for latent Dirichlet allocation

1) The full conditional distribution of $\boldsymbol{\gamma}_k$ is

$$(\boldsymbol{\gamma}_k|\boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{z}) \sim Dirichlet(\boldsymbol{N}_k + \boldsymbol{\beta}),$$

where $\boldsymbol{N}_k = (N_{k1}, \ldots, N_{kV})^T$.

2) The full conditional distribution of $\boldsymbol{\theta}_i$ is

$$\boldsymbol{\theta}_i \sim Dirichlet(\boldsymbol{N}_i + \boldsymbol{\alpha}),$$

where $\boldsymbol{N}_i = (N_{i1}, \ldots, N_{iK})^T$.

3) The full conditional distribution of $z_{i,n}$ is

$$\left(z_{i,n}|\boldsymbol{\theta}_i, w_{i,n} = v, \boldsymbol{\gamma}\right) \sim Multinomial(\theta_{i1}\gamma_{1v}, \theta_{i2}\gamma_{2v} \dots, \theta_{iK}\gamma_{Kv}).$$

C.3 Calculating DIC and AICM for Latent Dirichlet Allocation[5]

Suppose $\boldsymbol{\Theta}$ is a set of parameters that we focus on and $\boldsymbol{y}$ is observed data. DIC is defined as

$$DIC = D(\overline{\boldsymbol{\Theta}}) + 2p_D,$$

where

$$D(\boldsymbol{\Theta}) = -2\log\{p(y|\boldsymbol{\Theta})\} + 2\log\{f(y)\},$$

and

$$p_D = \overline{D(\boldsymbol{\Theta})} - D(\overline{\boldsymbol{\Theta}}),$$

which is the effective number of parameters. The second term of $D(\boldsymbol{\Theta})$, $2\log\{f(y)\}$, is a constant in that it is not influenced by model parameters. Thus, we ignore the second term here for model comparison purposes.

AICM is defined as

$$AICM = 2\left(\overline{l(\boldsymbol{\Theta})} - s^2_{l(\boldsymbol{\Theta})}\right),$$

where

$$\overline{l(\boldsymbol{\Theta})} = \frac{1}{S}\sum_{s=1}^{S} l\left(\boldsymbol{\Theta}^{(s)}\right)$$

and

[5] For the prior distributions of the LDA parameters, we chose $\boldsymbol{\alpha} = \mathbf{1}_K$ and $\boldsymbol{\beta} = \mathbf{1}_V$ so that $\boldsymbol{\theta}_{i,\cdot,t}$ and $\boldsymbol{\gamma}_k$ have uniform distribution on the simplex. Here, $\mathbf{1}_K$ indicates a $K \times 1$ vector consisted of 1.

$$s_l^2(\boldsymbol{\Theta}) = \frac{1}{S}\sum_{s=1}^{S}\left(l\left(\boldsymbol{\Theta}^{(s)}\right) - \overline{l(\boldsymbol{\Theta})}\right)^2.$$

Here, $l(\boldsymbol{\Theta}) = \log\{p(y|\boldsymbol{\Theta})\}$, $\boldsymbol{\Theta}^{(s)}$ is a sample from the posterior distribution at the sth iteration, and S is the total number of iterations.

DIC depends on the choice of focused parameters. If we focus on $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \boldsymbol{\gamma}\}$, $l(\boldsymbol{\Theta})$ and $\overline{l(\boldsymbol{\Theta})}$ for calculating DIC and AICM can be obtained as follows:

$$l(\boldsymbol{\Theta}) = \log\{p(\boldsymbol{w}|\boldsymbol{\theta}, \boldsymbol{\gamma})\} = \sum_{i=1}^{D}\sum_{v\in\boldsymbol{w}_i}\log\left(\sum_{k=1}^{K}\gamma_{kv}\theta_{ik}\right),$$

and

$$l(\bar{\boldsymbol{\Theta}}) = \sum_{i=1}^{D}\sum_{v\in\boldsymbol{w}_i}\log\left(\sum_{k=1}^{K}\bar{\gamma}_{kv}\bar{\theta}_{ik}\right),$$

where

$$\bar{\gamma}_{kv} = \frac{1}{S}\sum_{s=1}^{S}\gamma_{kv}^{(s)}, \qquad \bar{\theta}_{ik} = \frac{1}{S}\sum_{s=1}^{S}\theta_{ik}^{(s)}.$$

In addition, $\overline{l(\boldsymbol{\Theta})}$ can be obtained as

$$\overline{l(\boldsymbol{\Theta})} = \frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{D}\sum_{v\in\boldsymbol{w}_i}\log\left(\sum_{k=1}^{K}\gamma_{kv}^{(s)}\theta_{ik}^{(s)}\right).$$

C.4 Calculating DIC and AICM for MixSLDA

For MixSLDA, if we focus on parameters $\boldsymbol{\Theta} = \left\{\boldsymbol{\theta}, \boldsymbol{\gamma}, \widetilde{\boldsymbol{b}}, \widetilde{\boldsymbol{\sigma}}^2, \boldsymbol{\pi}\right\}$, $l(\boldsymbol{\Theta})$ and $\overline{l(\boldsymbol{\Theta})}$ for calculating DIC and AICM can be obtained as follows:

$$l(\boldsymbol{\Theta}) = \log\{p(\boldsymbol{w}, \boldsymbol{y}|\boldsymbol{\Theta})\} = \log\{p(\boldsymbol{w}|\boldsymbol{\theta}, \boldsymbol{\gamma})\} + \log\left\{p\left(\boldsymbol{y}|\boldsymbol{\theta}, \widetilde{\boldsymbol{b}}, \widetilde{\boldsymbol{\sigma}}^2, \boldsymbol{\pi}\right)\right\}$$

$$= \log\left\{\prod_{i=1}^{D}\prod_{n=1}^{N_i}p\left(w_{i,n}|\boldsymbol{\theta}_i, \boldsymbol{\gamma}\right)\right\} + \log\left\{\prod_{i=1}^{D}p\left(y_i|\boldsymbol{\theta}_i, \widetilde{\boldsymbol{b}}, \widetilde{\boldsymbol{\sigma}}^2, \boldsymbol{\pi}\right)\right\}$$

$$= \sum_{i=1}^{D}\sum_{n=1}^{N_i} \log\left\{\sum_{z_{n,i}} p\left(w_{i,n}, z_{i,n}|\boldsymbol{\theta}_i, \boldsymbol{\gamma}\right)\right\} + \sum_{i=1}^{D} \log\left\{\sum_{C_g} p\left(y_i, C_g|\boldsymbol{\theta}_i, \widetilde{\boldsymbol{b}}, \widetilde{\boldsymbol{\sigma}}^2, \boldsymbol{\pi}\right)\right\}$$

$$= \sum_{i=1}^{D}\sum_{n=1}^{N_i} \log\left\{\sum_{z_{n,i}} p\left(w_{i,n}|z_{i,n}, \boldsymbol{\gamma}\right)p\left(z_{i,n}|\boldsymbol{\theta}_i\right)\right\} + \sum_{i=1}^{D} \log\left\{\sum_{C_g} p\left(C_g|\boldsymbol{\pi}\right)p\left(y_i|C_g, \boldsymbol{\theta}_i, \widetilde{\boldsymbol{b}}, \widetilde{\boldsymbol{\sigma}}^2\right)\right\}$$

$$= \sum_{i=1}^{D}\sum_{v\in \boldsymbol{w}_i} \log\left(\sum_{k=1}^{K} \gamma_{kv}\theta_{ik}\right) + \sum_{i=1}^{D} \log\left\{\sum_{g=1}^{G} \pi_g \left[\frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\left\{-\frac{1}{2\sigma_g^2}\left(y_i - \boldsymbol{b}_g^T\boldsymbol{\theta}_i\right)^2\right\}\right]\right\},$$

and similarly,

$$l(\bar{\boldsymbol{\Theta}}) = \sum_{i=1}^{D}\sum_{v\in \boldsymbol{w}_i} \log\left(\sum_{k=1}^{K} \bar{\gamma}_{kv}\bar{\theta}_{ik}\right) + \sum_{i=1}^{D} \log\left\{\sum_{g=1}^{G} \bar{\pi}_g \left[\frac{1}{\sqrt{2\pi\bar{\sigma}_g^2}} \exp\left\{-\frac{1}{2\bar{\sigma}_g^2}\left(y_i - \bar{\boldsymbol{b}}_g^T\bar{\boldsymbol{\theta}}_i\right)^2\right\}\right]\right\},$$

where

$\bar{\gamma}_{kv} = \frac{1}{S}\sum_{s=1}^{S}\gamma_{kv}^{(s)}$, $\bar{\theta}_{ik} = \frac{1}{S}\sum_{s=1}^{S}\theta_{ik}^{(s)}$, $\bar{\boldsymbol{b}}_g = \frac{1}{S}\sum_{s=1}^{S}\boldsymbol{b}_g^{(s)}$, $\bar{\sigma}_g^2 = \frac{1}{S}\sum_{s=1}^{S}{\sigma^2}_g^{(s)}$, $\bar{\pi}_g = \frac{1}{S}\sum_{s=1}^{S}\pi_g^{(s)}$.

In addition, $\overline{l(\boldsymbol{\Theta})}$ can be obtained as

$$\overline{l(\boldsymbol{\Theta})} = \frac{1}{S}\sum_{s=1}^{S}\left[\sum_{i=1}^{D}\sum_{v\in \boldsymbol{w}_i} \log\left(\sum_{k=1}^{K} \gamma_{kv}^{(s)}\theta_{ik}^{(s)}\right)\right.$$

$$\left. + \sum_{i=1}^{D} \log\left\{\sum_{g=1}^{G} \pi_g^{(s)}\left[\frac{1}{\sqrt{2\pi{\sigma^2}_g^{(s)}}} \exp\left\{-\frac{1}{2{\sigma^2}_g^{(s)}}\left(y_i - \left(\boldsymbol{b}_g^{(s)}\right)^T\boldsymbol{\theta}_i^{(s)}\right)^2\right\}\right]\right\}\right].$$