# Exploring the use of Statistical Topic Models with Item Response Models in Educational Measurement

by

Jordan M. Wheeler

(Under the Direction of Allan S. Cohen)

Abstract

Topic models are statistical models used to analyze textual data. The objective of most topic models is to interpret the latent semantic space of a set of related texts. The use of topic models within the field of educational measurement has increased in recent years as a method for analyzing constructed-response (CR) items. These approaches typically utilize two different topic models: Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). Both of these topic models have shown to be useful in a variety of applications within educational measurement applications, however, there are limited studies that investigate how these models can be used with item response models. This dissertation, therefore, focuses on how topic models can be used with item response models to improve the quality and interpretation of measures. This focus is addressed using three studies. The first study compares LSA and LDA using a simulation study. The objective of this study is to understand how these models compare to each other in order to better understand the appropriate uses of each model. The second study uses LDA to analyze an unfolding scale. In this study, LDA is able to help define the latent unfolding scale, and is able to help define the inaccuracies of raters on constructed-response items. The third study proposes a new scoring procedure for mixed-format assessments that incorporates information obtained from topic models with an item response model to improve the accuracy of ability estimates.

Index words: topic models, mixed-format assessments, item response theory, latent semantic analysis, latent Dirichlet allocation

# Exploring the use of Statistical Topic Models with Item Response Models in Educational Measurement

by

Jordan M. Wheeler

B.S., Nebraska Wesleyan University, 2017
M.S., University of Nebraska - Omaha, 2019

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

2023

# Exploring the use of Statistical Topic Models with Item Response Models in Educational Measurement

by

Jordan M. Wheeler

Major Professor: Allan S. Cohen

Committee: George Engelhard Jr.
Shiyu Wang
Jue Wang

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
May 2023

# Dedication

*To my parents, Rosemary and Michael Wheeler, you are the ones that have shaped me into the person I have become. I am forever grateful for your unconditional love and support.*

# Acknowledgments

My graduate experience and this dissertation would not have been successful without the help of so many incredible people in my life. First, I would like to thank my advisor, Allan Cohen, whose profound knowledge and passion for quantitative methods and psychometrics proved invaluable to me. Your mentorship has transformed me into a better researcher, writer, presenter, and person. I am immensely grateful for your unwavering patience and encouragement, and I look forward to our continued collaboration.

I am equally grateful to my committee members and other mentors who have helped shape my graduate journey. Professor Engelhard, you have always been incredibly patient and encouraging during graduate school. Our collaborations have nurtured my research, writing, and mentoring skills, and I am proud to have worked with you. Our collaborations have helped me become a better researcher and writer. Dr. Shiyu Wang, your statistical knowledge and creativity have been instrumental in my development as a researcher. I am thankful for the opportunities I have had to collaborate with you and am excited for our future projects. Dr. Jue Wang, I am thankful for our collaboration and look forward to working with you more in the future. Laine Bradshaw, I will be forever grateful for the opportunity you have given me to work with Navvy. Your expertise and passion for measurement are inspiring.

I also want to thank the other remarkable professors I met at the University of Georgia, including Dr. Matt Madison, Dr. Seock-Ho Kim, and Dr. Larua Lu, for their encouragement and guidance.

I would also like to thank some of my closest friends, Tyler, Tyler, and Brad, for making this experience more enjoyable. I am so grateful for the Zoom happy hours and times you visited me in Athens. I am also grateful for the friends I made throughout my time at Georgia, especially Jake and Katie, and to my fellow classmates, Jiawei, Madeline, Elaine, Zack, Ye, Cony, and others.

I want to thank my parents. You have instilled resilience and a strong work ethic within me. Without those skills, this journey would not have been possible. I would also like to thank my in-laws, Sue and Lee. You have always been supportive and encouraging. I appreciate all you have done for Carly and me during this process. I am grateful for my siblings, nieces, and in-laws. Life is much better with a great family.

Lastly (and most importantly), I want to express my deepest gratitude to my incredible wife, Carly. Throughout my graduate school journey, you have selflessly supported, encouraged, and motivated me. You have been by my side through the best and worst parts of this experience. I am incredibly thankful for your sacrifices, especially your willingness to leave your friends and family to come to Athens with me. This journey could have never happened without you. I love you.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction and Literature Review

Topic models are mathematical and statistical models used to analyze sets of documents (i.e., corpus) and were initially developed as an indexing technique to summarize, compare, and group large sets of textual data. The primary objective of topic models is to discover topics and the usage of topics within a corpus (Rosen-Zvi et al., 2010). Topics are textual features that occur across all documents and can often be thought of as latent clusters of words. The usage of topics are textual features that occur within an individual document and express the use of topics within the document.

Although topic models were originally developed for the information retrieval and machine learning fields, they have a growing interest within educational measurement as a technique to analyze constructed-response (CR) items. Two commonly used topic models within educational measurement are latent semantic analysis (LSA; Deerwester et al., 1990) and latent Dirichlet allocation (LDA; D. M. Blei et al., 2003). LSA is a matrix-algebra technique similar to principal component analysis and factor analysis. It uses a matrix factorization technique known as singular value decomposition to extract latent topics from a set of

documents. LDA is a statistical model that uses an assumed generative model to estimate latent topics.

## 1.1 Applications of Topic Models in Educational Measurement

There are a variety of applications and purposes of topic models within educational measurement, including automated scoring algorithms for CR items, a method for performing simple qualitative analyses, assessing students' textual responses, identifying misconceptions, and detecting complex forms of cheating (e.g., Foltz et al., 1999; S. Kim et al., 2017; Mozgovoy et al., 2010; Shin et al., 2019).

One of the first uses of topic models in educational measurement was for automated scoring of CR items. Many automated scoring algorithms utilize topic models due to their ability to produce scores that are consistent with human raters. For example, Intelligent Essay Assessor (IEA; Foltz et al., 1998), e-rater (Attali & Burstein, 2006), and Coh-Metrix (Graesser et al., 2004) are well-established automated scoring software that use a variety of methods including LSA. Specifically, these automated score software packages extract text features using LSA and use the features as covariates in a scoring algorithm, such as a regression model, decision trees, or neural-networks. LDA has also been used for automated scoring algorithms, however, the performance of these algorithms is consistently worse than LSA (Kakkonen et al., 2006; Xiong et al., 2021).

Beyond scoring CR items, topic models have been used to analyze students' textual responses in order to better understand students' thinking, reasoning, and strategies while responding to CR items. For example, Choi et al. (2017) used LDA to detect latent topics to aid in identifying the semantic and thematic structure of students' writing. Topics that differentiated the types of words and the latent themes in good versus poor writing were

detected. S. Kim et al. (2017) found that two of the three topics in the topic model of answers to a test of science inquiry knowledge corresponded to academic language and discipline-specific language as reflected in the use of scientific words such as measure, effect, and hypothesis. Answers making use primarily of the third topic used mainly everyday language. Wheeler, Raczynski, et al. (2022) used LDA to assess responses to an integrated writing assessment and found different types of writing strategies that students used. The results of that study showed that LDA illuminated students' reasoning when responding to CR items and showed how it is associated with their scores.

Topic models can also be used to assess learning and growth in students' writing. For example, Duong et al. (2019) used LDA to analyze answers to an essay test requiring integrative borrowing. Students' responses taught to use instructional conversation (a form of collaborative learning) were compared to students who had not. Results indicated that the answers of students taught instructional conversation expanded beyond the characters and settings in the prompt and expressed an understanding of multiple perspectives than the answers of students who had not had an instructional conversation. Students who had not had instructional conversations simply re-stated the prompt did not change the setting and only took a single perspective into account. Kwak (2019) reported a change in the use of topics detected by LDA in essay test answers due to instruction.

In addition to scoring, measuring growth, and assessing students' thinking and reasoning, topic models have been useful for different applications. Shin et al. (2019) used LDA to analyze written responses to CR items to identify topics that were associated with common misconceptions. In that study, they used this information to develop distractors for multiple-choice (MC) items. S. Kim et al. (2017) showed that the topic structure from LDA models yielded similar results to a qualitative analysis of the same essays. Mozgovoy et al. (2010) used LSA to detect cheating on CR items and mixed-format assessments.

## 1.2 Connection between Topic Models and Item Response Models

Item response models (IRM) are statistical models that are used to analyze students' responses to items on an assessment. The objective of IRMs is to connect students and items on the same latent trait (i.e., ability) using an assumed model. In this regard, IRMs estimate the ability of students and the difficulty of items. Specifically, IRMs assume that the probability of a student correctly responding to an item is a function of the student's ability and item's difficulty.

In practice, IRMs are applied to scored item response in order to obtain an estimate of the ability of each student. The scored item responses for MC items are typically determined from a key and the scored item response for CR items are typically determined from a rubric. Currently, one of the main connections between topic models and IRMs is through automated scoring algorithms. That is, topic models are sometimes used as automated scoring algorithms to provide rubric scores to CR items.

## 1.3 Purpose of Study

Although topic models have a variety of uses within educational measurement, there is a distinction between the uses of LSA and LDA. More specifically, LSA is often only used in educational measurement as a method for automated scoring algorithms of CR items whereas LDA is often only used in educational measurement for supplement analyses of CR items in order to provide substantive interpretations of the topics used by students. This is particularly noteworthy because these topic models have the same goals (i.e., find latent topics) and use the same data. Moreover, there have been limited studies that have investigated the differences between these two models to better understand why this distinction exists.

Despite the distinct aims of LSA and LDA, there has been limited research on their potential integration with item response models beyond mere rubric-based scoring of constructed-response (CR) items. In reality, topic models can offer much more than automated scoring algorithms, but their use has been limited. Given the increase of computer-based assessments, however, incorporating topic models into item response models has become more feasible and could prove highly beneficial.

This dissertation focuses on the use of topic models within educational measurement. Furthermore, the objective of this study is to investigate and address the differences between LSA and LDA, and how these methods can be integrated and used with item response models to improve the quality and interpretation of measures. Specifically, this dissertation investigates the following research questions:

1. How do commonly used topic models within educational measurement differ?
2. Can topic models be used to help interpret a latent scale obtained from an item response model?
3. Can the information obtained from topic models be used to increase the validity and reliability of IRT ability estimates for test takers?

## 1.4 Overview of Dissertation

The dissertation contains three studies. The first study compares LSA and LDA using a simulation study. The objective of this study is to understand how these models compare to each other in order to better understand the appropriate uses of each model. The second study uses LDA to analyze an unfolding scale. In this study, LDA is able to help define the latent unfolding scale, and is able to help define the inaccuracies of raters on constructed-response items. The third study proposes a new scoring procedure for mixed-format assessments that

incorporates information obtained from topic models with an item response model to improve the accuracy of ability estimates.

# Chapter 2

# A Comparison of Latent Semantic Analysis and Latent Dirichlet Allocation in Educational Measurement[1]

[1] Wheeler, J. M., Cohen, A. S., & Wang, S. Submitted to *Journal of Educational and Behavioral Statistics*, 2/23/2023

## 2.1 Preface

The study presented in this chapter is currently under revision in Wheeler, Cohen, and Wang (under revision). The study was directed under the supervision of Drs. Shiyu Wang and Allan S. Cohen. My role in this study was to lead the work for the comparison of LSA and LDA, develop the methodology used to compare the semantic spaces obtained from both models, develop the R code that was used to run the simulation study, and draft the initial manuscript. This paper highlights the foundational concepts for two commonly used topic models, LSA and LDA, within an educational measurement context. It also introduces a new methodology for measuring the similarity between the results obtained by these two topic models.

## 2.2 Abstract

Topic models are mathematical and statistical models used to analyze textual data. The objective of topic models is to gain information about the latent semantic space of a set of related textual data. The semantic space of a set of textual data contains the relationship between documents and words and how they are used. Topic models are becoming more common in educational measurement research as a method for analyzing students' responses to constructed-response (CR) items. Two popular topic models are latent semantic analysis (LSA) and Latent Dirichlet Allocation (LDA). LSA uses linear algebra techniques whereas LDA uses an assumed statistical model and generative process. In educational measurement, LSA is often used in algorithmic scoring of essays due to its high reliability and agreement with human raters, and LDA is often used as a supplemental analysis to gain additional information about students, such as their thinking and reasoning. In this paper, we review and compare the LSA and LDA topic models. We also introduce a methodology for comparing

the semantic spaces obtained by the two models and use a simulation study to investigate their similarities.

## 2.3 Introduction

Topic models are a family of statistical models and mathematical techniques used to analyze textual data. The objective of topic models is to gain information about the latent semantic space of a set of related textual data. The semantic space contains the relationship between the textual data and is obtained through estimating latent topics (sometimes referred to as latent components or features). The latent topics discovered by topic models can be thought of as a cluster of words that illuminate common themes within the set of textual data. Topic models were initially developed for the machine learning and information retrieval fields, however, some topic models have recently gained interest within educational measurement as a technique for scoring and analyzing constructed-response (CR) items and mixed-format assessments (e.g., Shermis & Burstein, 2013; Wheeler, Wang, Tan, & Cohen, 2022). Two commonly used topic models in educational measurement are latent semantic analysis (LSA; Deerwester et al., 1990; Foltz et al., 1998; Landauer et al., 1998) and latent Dirichlet allocation (LDA; D. M. Blei et al., 2003).

In educational measurement, the LSA and LDA topic models are typically used to analyze a set of responses to a CR item. The purpose of these two models is to analyze the latent semantic space, which is the mathematical representation of the common topics used when responding to the CR item and describe the relationship between the responses. The applications of these two models, however, are different even though they use the same data. Specifically, LSA is often used as the basis for automated scoring engines of CR items, such as the commercially used Intelligent Essay Assessor (IEA; Foltz et al., 1999). Utilizing LSA for automated scoring greatly reduces the cost and time to score large numbers of CR items, and it has been shown to produce similar scores as human raters (Dikli, 2006; Shermis &

Burstein, 2013). On the other hand, LDA is often used to analyze the content of CR items to gain additional information beyond a score. LDA has multiple applications including identifying misconceptions, assessing the fairness of human raters, and analyzing the thinking and reasoning of students (e.g., Shin et al., 2019; Wheeler, Engelhard, et al., 2022a; Wheeler, Raczynski, et al., 2022). Although LDA has many applications, it has been shown to perform worse than LSA when it comes to automated scoring of CR items (Choi et al., 2017).

Even though LSA and LDA have different applications, these two models are trying to capture similar information about the responses to a CR item. That is, both models try to estimate the latent semantic space. There are limited studies, however, that investigate why these two models have different uses even though they use the same data. In this regard, it is important to investigate and compare the latent semantic spaces obtained from both models to better understand why the results have different uses and how they can be augmented to better improve the analysis of CR items.

### 2.3.1 Purpose of Study

The purpose of the study is twofold. First, since these topic models have increasingly been used in educational measurement, this study provides a thorough introduction to both topic modeling methods and their applications. Second, the measurement literature suggests that these two methods should be used for different purposes even though both methods use the same textual data to estimate semantic structures. It is important to compare the semantic spaces obtained from these two models in order to better understand why these models are useful for different applications and how these two models can be used together to provide better information about students' responses. The investigation and comparison of these two methods lead to an understanding of how these methods can be used to gain a more complete analysis of students' writing.

This paper is structured as follows: Section 2.4 presents the LSA and LDA topic models, along with their respective mathematical procedures. Section 2.5 introduces a method for comparing the latent semantic spaces obtained from LSA and LDA. In section 2.6, we conduct a simulation study that compares the semantic spaces of LSA and LDA under multiple conditions. Section 2.7 uses real data for an empirical comparison between LSA and LDA. Finally, in Section 2.8, we discuss the application of these models in educational measurement and highlight the similarities and differences between the two topic models.

## 2.4 Topic Models

This section introduces the terminology of topic models and a general notation for mathematically representing textual data. In addition to terminology and notation, the LSA and LDA topic models are presented and described.

### 2.4.1 Topic Model Notation

Before applying a topic model, the textual data needs to be converted into mathematical notation. In this subsection, we introduce several definitions and notations that are commonly used for topic models. In topic models, a *word* is a unit of discrete data, a *document* is a set of words, and a *corpus* is a set of documents. In educational measurement, a document is typically a single student response to an item and a corpus is the set of student responses to the same item. The *vocabulary* is the set of unique words observed in the entire corpus, and words are represented mathematically by their position in the vocabulary.

Suppose a corpus contains $D$ documents and has a vocabulary of $V$ words. Document $d \in \{1, ..., D\}$ contains $N_d$ words and is denoted by $w_d = (w_1, ..., w_{N_d})$ where $w_n$ is the $n$th word in the document and is represented by the position of the word in the vocabulary, such that $w_n \in \{1, ..., V\} \; \forall \; n = 1, ..., N_d$.

### 2.4.2 Latent Semantic Analysis

LSA is a matrix-based topic model that utilizes a matrix decomposition method known as singular value decomposition (SVD; Golub & Reinsch, 1971). LSA uses SVD to extract latent topics from a corpus. The latent topics are often referred to as the latent semantic space of a corpus, which is the mathematical representation for the meaning of words and documents. SVD extracts the latent topics by using the observed variation between words and documents. Once SVD is applied to a corpus, LSA selects a subset of the latent topics that explain the majority of the observed variation in documents. The subset of latent topics is used to mathematically represent the corpus (Landauer et al., 1998). The LSA process is similar to factor analysis and principal component analysis in that it searches for observed variations in response data with respect to latent factors and uses a subset of the latent factors to derive meaning from the response data (Tipping & Bishop, 1999; Wall et al., 2003).

#### Document-by-Word Matrix

LSA constructs a document-by-word matrix of the corpus by counting the number of times each word in the vocabulary appears in each document. The document-by-word matrix is defined as

$$\boldsymbol{X}_{D\times V} = [x_{d,v}], \tag{2.1}$$

where $x_{d,v}$ is the number of times word $v \in \{1, ..., V\}$ from the vocabulary occurs in document $d \in \{1, ..., D\}$ such that

$$x_{d,v} = \sum_{w_n \in \boldsymbol{w}_d} 1_{\{w_n = v\}}. \tag{2.2}$$

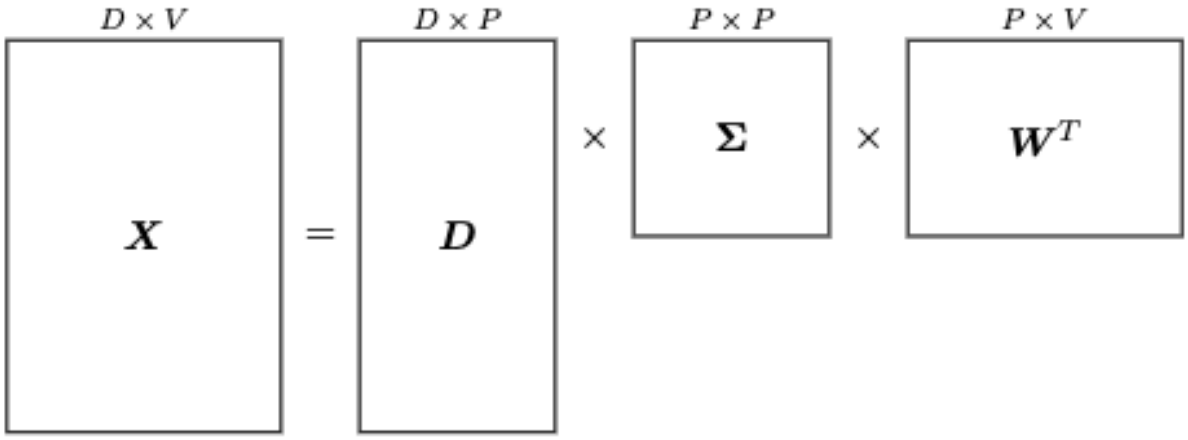
### Singular Value Decomposition

The document-by-word matrix, $\boldsymbol{X}_{D\times V}$, can be factored into three matrices using SVD, as shown in Figure 2.1, and can be mathematically defined as (Wild, 2016, see p.117 for more details):

$$\boldsymbol{X}_{D\times V} = \boldsymbol{D}_{D\times P}\boldsymbol{\Sigma}_{P\times P}\boldsymbol{W}^T_{P\times V}, \tag{2.3}$$

where $P = rank(\boldsymbol{X}_{D\times V})$ is the number of linearly independent eigenvectors (i.e., topics) extracted from the document-by-word matrix; $\boldsymbol{D}_{D\times P} = eigenvectors(\boldsymbol{X}\boldsymbol{X}^T)$ is the document matrix and contains the right-singular eigenvectors; $\boldsymbol{W}_{V\times P} = eigenvectors(\boldsymbol{X}^T\boldsymbol{X})$ is the word matrix and contains the left-singular eigenvectors; $\boldsymbol{\Sigma}_{P\times P} = diag(\sqrt{\lambda_1}, ..., \sqrt{\lambda_P}) = diag(\sigma_1, ..., \sigma_P)$, where $\sigma_1 \geq ... \geq \sigma_P$ is the diagonal topic matrix and contains the singular values (i.e., the square-roots of the eigenvalues) of the left- and right-singular eigenvectors.

Figure 2.1: Graphical representation of the Singular Value Decomposition model on the document-by-word matrix



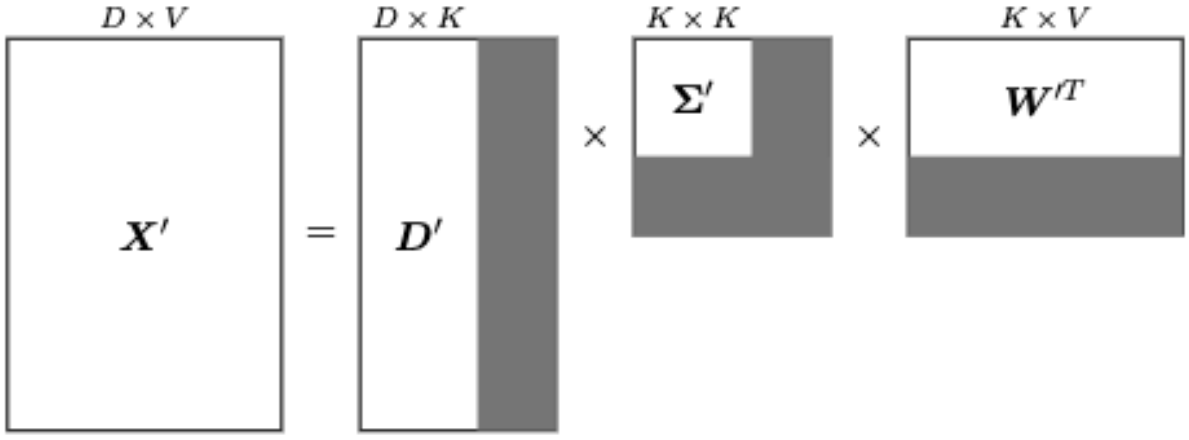

In the context of LSA and textual data, the matrices derived from SVD contain useful information about the underlying latent semantic space of the corpus. Specifically, $\boldsymbol{D}_{D\times P}$ contains information about the relationship between the individual documents and the latent

topics by representing how much of each topic is elicited in each document. $\boldsymbol{W}_{V\times P}$ contains information about the relationship between the words in the vocabulary and the latent topics by representing the importance of each word in each topic. $\boldsymbol{\Sigma}_{P\times P}$ contains the relative weights for the latent topics and provides information about the amount of variation that can be explained by each topic (Steinberger, Jezek, et al., 2004).

### LSA Dimensionality Reduction

Once the document-by-word matrix is factorized, the dimensionality, $P$, of the underlying latent semantic space is reduced by removing the latent topics with small singular values from the three matrices obtained from SVD (Landauer et al., 1998). Specifically, topics with relatively small singular values are removed from the topic matrix and their corresponding eigenvectors are removed from the document and word matrices. This is an important step in LSA since the initial dimensionality is typically large and often a large portion of the latent topics explain a small amount of the variation in the corpus. Additionally, the reduction results in a more useful representation of the semantic space and provides a clearer and more reliable relationship between documents and words (Deerwester et al., 1990).

Figure 2.2: Graphical representation of Latent Semantic Analysis (LSA) dimensionality reduction of a document-by-word matrix



Following the factorization in Equation (2.3), the dimensionality of the three matrices is reduced to $K$ latent topics where $0 \leq K \leq P$. Therefore, the $K+1, ..., P$ singular values and their corresponding eigenvectors are removed from the three matrices. The reduction results in a reduced document matrix, $\boldsymbol{D}'_{D\times K}$, a reduced word matrix, $\boldsymbol{W}'_{V\times K}$, and a reduced topic matrix, $\boldsymbol{\Sigma}'_{K\times K} = diag(\sigma_1, ..., \sigma_K)$ where $\sigma_1 \geq ... \geq \sigma_K$. As shown in Figure 2.2, the resulting reduced matrices can be multiplied to reconstruct the document-by-word matrix, with the original $D \times V$ dimensions, on a reduced latent semantic space:

$$\boldsymbol{X}'_{D\times V} = \boldsymbol{D}'_{D\times K}\boldsymbol{\Sigma}'_{K\times K}\boldsymbol{W}'^{T}_{K\times V}, \tag{2.4}$$

where $\boldsymbol{X}'_{D\times V}$ approximates $\boldsymbol{X}_{D\times V}$ and is a projection of the original document-by-word matrix on $K$ topics. It is important to note that $\boldsymbol{X}_{D\times V}$ is typically sparse (contains mostly 0s) because most documents only use a small subset of the words from the vocabulary (Wild, 2016, see p. 81). The resulting $\boldsymbol{X}'_{D\times V}$ is no longer sparse and better reflects the semantic space of the corpus (Deerwester et al., 1990; Wild, 2016).

### Model Selection

There are multiple methods for selecting an appropriate number of $K$ topics. In the context of using LSA for machine scoring of constructed-response items on assessments, $K$ can be determined by the performance of the scoring algorithm. Landauer and Dumais (1997) and Landauer et al. (1998) selected $K$ by maximizing the accuracy between scores obtained from human raters and the automated scoring engine. Shermis and Burstein (2013) suggested selecting $K$ that maximizes several metrics, including classification accuracy and Cohen's Kappa. More generally, these selection methods for $K$ are dependent on an operational criterion of a downstream task.

More recently, Wild (2016) developed a selection method for $K$ that is grounded in mathematics rather than operational criteria (Wild, 2016, see p. 123 for more details). The

method involves calculating a cutoff, $c$, for the sum of eigenvalues obtained from SVD, given by

$$c = 0.8 \times Tr(\boldsymbol{X}\boldsymbol{X}^T), \tag{2.5}$$

where $Tr()$ is the trace function of a matrix. In other words, the cutoff value is determined by taking 80% of the trace from the document-by-word matrix multiplied by its transpose. The number of topics, $K$, is determined by the minimum number of singular values needed for the sum of their squares to be greater than the cutoff value, such that

$$K = \min\{k | \sum_{p=1}^{k} \sigma_p^2 \geq c\}. \tag{2.6}$$

This method substantially reduces the computational cost and time of LSA by iteratively calculating the eigenvectors and eigenvalues for SVD until the cutoff condition is met.

### Application of LSA

Educational measurement research utilizes LSA in a variety of ways. One application of LSA is to compare the similarity between documents or between words. This type of application often calculates a vector correlation, such as the cosine similarity measure, between two rows (for documents) or two columns (for words) in $\boldsymbol{X}'_{D\times V}$ (Deerwester et al., 1990; Landauer & Dumais, 1997). The cosine similarity measure calculates the angle between two vectors of the same dimension. In the context of LSA, the cosine similarity measure expresses how similar two documents or words are in the semantic space of the corpus. Another application is using LSA to extract features from the documents that are then used in statistical or machine learning models (e.g., Graesser et al., 2004; LaVoie et al., 2020; Wheeler, Wang, Tan, & Cohen, 2022). The set of features for the documents are calculated by multiplying the reduced document matrix by the weights of the topic matrix, $\boldsymbol{D}'_{D\times K}\boldsymbol{\Sigma}'_{K\times K}$. LSA can

also be used to cluster documents and words into similar groups (Wild, 2016). The clusters of documents and words can be visually represented in a 2D- or 3D-graph by plotting the first two or three points in the reduced document and word matrices (Wild, 2016, see Fig. 4.8 on p. 89).

LSA can also be applied to a set of $M$ unseen documents. The set of unseen documents is used to construct a new document-by-word, $\boldsymbol{X}_{\text{new}(M\times V)}$, following Equations (1) and (2). Note that the new document-by-word matrix uses the same vocabulary as the original document-by-word matrix even though the new set of documents may have additional words not seen in the original document-by-word matrix. New words are ignored since LSA is dependent on words used to create the semantic space; that is, the eigenvectors calculated are dependent on the columns of the original document-by-word matrix. This is often seen as a limitation of LSA.

Once the new document-by-word matrix is constructed, the matrix can be projected onto the semantic space derived from the original document-by-word matrix using the following equation:

$$\boldsymbol{D}'_{new(M\times K)} = \boldsymbol{X}_{new(M\times V)}\boldsymbol{W}'_{V\times K}\boldsymbol{\Sigma}'_{K\times K}, \tag{2.7}$$

where $\boldsymbol{W}'_{V\times K}$ and $\boldsymbol{\Sigma}'_{K\times K}$ are the reduced word and topic matrices obtained from the original document-by-word matrix; and $\boldsymbol{D}'_{\text{new}(M\times K)}$ is the document matrix for the new, unseen documents.

### 2.4.3 Latent Dirichlet Allocation

LDA is a probabilistic mixed membership topic model that uses an assumed joint probability distribution and generative process to estimate the latent semantic space for a corpus. LDA is often used as an exploratory method for extracting latent features. In this regard, the LDA process fits multiple candidate models, with varying numbers of latent features. The candidate

models are compared using model fit indices, along with interpretability, to determine the number of latent features that best explain the corpus (Mardones-Segovia et al., 2021).

### LDA Parameters and Assumptions

There are three sets of parameters that the LDA model estimates: topics, topic proportions, and topic assignments. *Topics* are corpus-wide parameters which are essentially clusters of words that are commonly used by documents in the corpus. Topic proportions are document-wide parameters that express how much of each topic is used by each document. *Topic assignments* are parameters for each word that indicates the topic membership of each word in a document.

LDA assumes a priori that there are $K$ topics used in the corpus where each topic is a set of probabilities for every word in the vocabulary. Topic $k \in \{1, ..., K\}$ is assumed to follow a Dirichlet distribution over the $V$ words in the vocabulary and is denoted by $\boldsymbol{\beta}_k = (\beta_1, ..., \beta_V)$, where $\beta_v$ is the probability of the $v$th word in the vocabulary being used in the topic, such that $0 \leq \beta_v \leq 1 \; \forall \; v \in \{1, ..., V\}$ and $\sum_{v=1}^{V} \beta_v = 1$. Topic proportions for document $d \in \{1, ..., D\}$ are assumed to follow a Dirichlet distribution over the $K$ topics and are denoted by $\boldsymbol{\tau}_d = (\tau_1, ..., \tau_K)$, where $\tau_k$ is the proportion of the document that used the $k$th topic, such that $0 \leq \tau_k \leq 1 \; \forall \; k \in \{1, ..., K\}$ and $\sum_{k=1}^{K} \tau_k = 1$. Topic assignments for document $d \in \{1, ..., D\}$ is a set of discrete data and is denoted by $\boldsymbol{t}_d = (t_1, ..., t_{N_d})$, where $t_n$ is the topic assignment of the $n$th word in the document and is represented by a topic indicator, such that $t_n \in \{1, ..., K\} \; \forall \; n = 1, ..., N_d$.

### LDA Generative Process

LDA assumes a generative process for how all documents in the corpus are generated. This process assumes that a document is generated using the parameters described above, which is an important assumption for estimating the model parameters from the observed data.

The LDA generative model assumes that a corpus is generated through the following process (D. Blei et al., 2010; D. M. Blei et al., 2003):

A corpus consists of $K$ topics (assumed a prior) and $D$ documents. Each document $d \in \{1, ...D\}$ is generated as follows

1. Choose the length of the document, $N_d \sim Poisson(\lambda)$
2. Choose the topic proportions of the document, $\boldsymbol{\tau}_d \sim Dirichlet(\boldsymbol{\eta})$
3. For each word in the document, indexed as $n \in \{1, ...,_d\}$:
   (a) Choose topic assignment $t_n \sim Multinomial(\boldsymbol{\tau}_d)$
   (b) Given the topic assignment, choose word $w_n \sim Multinomial(\boldsymbol{\beta}_{k=t_n})$

**LDA Probabilistic Model**

LDA models the observed words, topic parameters, topic proportion parameters, and topic assignment parameters using the following joint distribution:

$$p(\boldsymbol{w}_{1:D}, \boldsymbol{t}_{1:D}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\tau}_{1:D}), \tag{2.8}$$

where $\boldsymbol{w}_{1:D} = \{\boldsymbol{w}_1, ..., \boldsymbol{w}_D\}$ is the set of documents in the corpus; $\boldsymbol{t}_{1:D} = \{\boldsymbol{t}_1, ..., \boldsymbol{t}_D\}$ is the set of topic assignments; $\boldsymbol{\beta}_{1:K} = \{\beta_1, ..., \beta_K\}$ is the set of topics; and $\boldsymbol{\tau}_{1:D} = \{\boldsymbol{\tau}_1, ..., \boldsymbol{\tau}_D\}$ is the set of topic proportions. Using the assumed generative process described above, the joint distribution in Equation (2.8) can be factorized as:

$$\begin{aligned}
&p(\boldsymbol{w}_{1:D}, \boldsymbol{t}_{1:D}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\tau}_{1:D}) = p(\boldsymbol{w}_{1:D}|\boldsymbol{t}_{1:D}, \boldsymbol{\beta}_{1:K}) p(\boldsymbol{t}_{1:D}|\boldsymbol{\tau}_{1:D}) p(\boldsymbol{\beta}_{1:K}) p(\boldsymbol{\tau}_{1:D}) \\
&= \prod_{k=1}^{K} p(\boldsymbol{\beta}_k|\boldsymbol{\nu}) \prod_{d=1}^{D} p(\boldsymbol{\tau}_d|\boldsymbol{\eta}) p(\boldsymbol{w}_d|\boldsymbol{t}_d, \boldsymbol{\beta}_{1:K}) p(\boldsymbol{t}_d|\boldsymbol{\tau}_d) \\
&= \prod_{k=1}^{K} p(\boldsymbol{\beta}_k|\boldsymbol{\nu}) \prod_{d=1}^{D} \left( p(\boldsymbol{\tau}_d|\boldsymbol{\eta}) \prod_{n=1}^{N_d} \left( p(w_{d,n}|t_{d,n}, \boldsymbol{\beta}_{k=t_{d,n}}) p(t_{d,n}|\boldsymbol{\tau}_d) \right) \right)
\end{aligned} \tag{2.9}$$

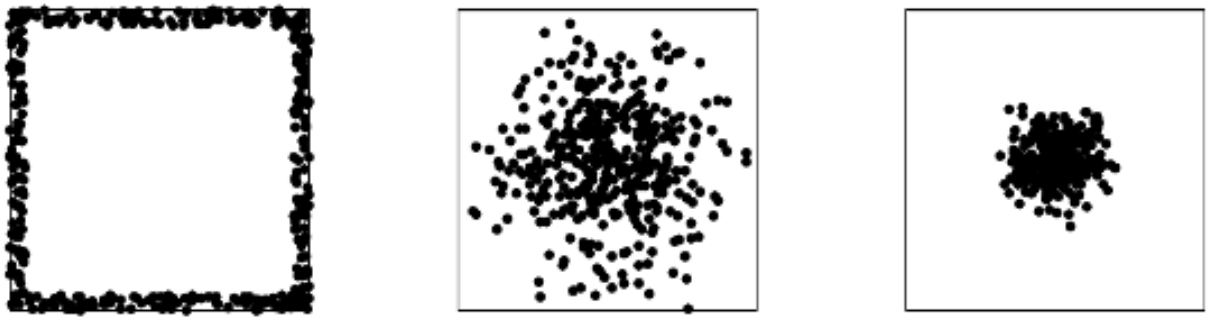
where $p(w_{d,n}|t_{d,n}, \beta_{k=t_{d,n}})$ is the conditional probability distribution of the nth word in the $d$th document given the topic assignment of the word and the topics, and is assumed to be a Multinomial distribution; $p(t_{d,n}|\boldsymbol{\tau}_d)$ is the conditional probability distribution topic assignment of the nth word in the $d$th document, and is assumed to be a Multinomial distribution; $p(\boldsymbol{\beta}_k|\boldsymbol{\nu})$ and $p(\boldsymbol{\tau}_d|\boldsymbol{\eta})$ are the probability distributions for the $k$th topic and the topic proportions of the $d$th document, respectively, and are assumed to be a Dirichlet distribution such that

$$p(\boldsymbol{\beta}_k|\boldsymbol{\nu}) = \frac{\Gamma(\sum_{v=1}^{V} \nu_v)}{\sum_{v=1}^{V} \Gamma(\nu_v)} \prod_{v=1}^{V} \beta_v^{\nu_v - 1}, \tag{2.10}$$

$$p(\boldsymbol{\tau}_d|\boldsymbol{\eta}) = \frac{\Gamma(\sum_{k=1}^{K} \eta_k)}{\sum_{k=1}^{K} \Gamma(\eta_k)} \prod_{k=1}^{K} \tau_k^{\eta_k - 1}, \tag{2.11}$$

where $\boldsymbol{\nu} = (\nu_1, ..., \nu_V)$ is the hyperparameter for the topic distributions, which are typically symmetric such that $\nu_1 = ... = \nu_V$; $\boldsymbol{\eta} = (\eta_1, ..., \eta_K)$ is the hyperparameter for the topic proportions and are also typically symmetric such that $\eta_1 = ... = \eta_K$; and $\Gamma$ is the standard gamma function, $\Gamma(\eta) = (\eta - 1)!$

Figure 2.3: Effects of the prior hyperparameters on topic proportions and word probability distributions



$\eta_1, \eta_2, ..., \eta_K < 1$
$\nu_1, \nu_2, ..., \nu_V < 1$

$\eta_1, \eta_2, ..., \eta_K = 1$
$\nu_1, \nu_2, ..., \nu_V = 1$

$\eta_1, \eta_2, ..., \eta_K > 1$
$\nu_1, \nu_2, ..., \nu_V > 1$

The values of the hyperparameters impact the shape of the Dirichlet distributions. Figure 2.3 illustrates the effects of the hyperparameters on the topic proportions and word probability distributions in the LDA model. The four sides of the square represent four individual topics while the dots in the square represent either documents or words. When the hyperparameter values are small, the probability mass of the Dirichlet distribution shifts to a few elements (i.e., a few words for topic distributions and a few topics for topic proportions). When the hyperparameter values are large, the probability mass of the Dirichlet distributions shifts to a uniform spread over all the elements (i.e., topics have a uniform probability overall for all words and all topic proportions have a uniform probability over all topics).

### LDA Parameter Inference

The joint distribution shown in Equation (2.8) is intractable due to its dimensionality (there are $K \times V + D \times K + \sum_{d=1}^{D} N_d$ parameters). The latent parameters (topics, topic proportions, and topic assignments) of LDA, therefore, are typically estimated through Bayesian techniques that infer the topic, topic proportion, and topic assignment parameters using the observed documents and the assumed joint probability distribution shown in Equation (2.9). Following Bayes' theorem, the probability distribution of the latent parameters given the observed documents can be expressed as

$$\begin{aligned} p(\boldsymbol{\tau}_{1:D}, \boldsymbol{\beta}_{1:K}, \boldsymbol{t}_{1:D} | \boldsymbol{w}_{1:D}) &= \frac{p(\boldsymbol{w}_{1:D}, \boldsymbol{t}_{1:D}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\tau}_{1:D})}{p(\boldsymbol{w}_{1:D})} \\ &\propto p(\boldsymbol{w}_{1:D}, \boldsymbol{t}_{1:D}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\tau}_{1:D}), \end{aligned} \tag{2.12}$$

where $p(\boldsymbol{\tau}_{1:D}, \boldsymbol{\beta}_{1:K}, \boldsymbol{t}_{1:D} | \boldsymbol{w}_{1:D})$ is referred to as the posterior probability distribution; and $p(\boldsymbol{w}_{1:D}, \boldsymbol{t}_{1:D}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\tau}_{1:D})$ is the joint distribution shown in Equations (7) and (8).

There are a variety of Bayesian estimation techniques to estimate the parameters, however, the two most commonly used techniques are variational inference using expectation-maximization and Gibbs sampling. Variational inference is an efficient estimation method,

but it is more difficult to implement and can get stuck at local maxima (D. M. Blei et al., 2003). Gibbs sampling algorithms are intuitive and easy to implement, however, the estimation can be slow and difficult to assess convergence. Griffiths and Steyvers (2004) provided a collapsed Gibbs sampling algorithm for LDA that integrates out the topics and topic proportions from the posterior and only estimates the topic assignments of each word. The topics and topic proportions are analytically computed from the estimated topic assignments.

Mardones-Segovia et al. (2022) conducted a simulation study that investigated the performance of the three estimation algorithms previously mentioned. The study used conditions that are often found in educational measurement settings (e.g., a small number of documents, a short document length, and a small number of topics). The results of the study found that in most conditions the Gibbs sampling algorithm recovered the LDA parameters more accurately than variational inference. That study suggests that Gibbs sampling algorithms for educational settings should be used over variation inference.

### Model Selection

Since the number of topics is assumed a priori, LDA is considered an exploratory model. That is, there are no mathematical methods to determine the correct number of topics in a corpus. An appropriate number of topics, therefore, needs to be determined through model selection indices. The process of selecting an appropriate number of topics is to estimate multiple candidate models and compare all models using a variety of model selection indices. The original LDA paper suggests selecting the number of topics using either an operational criterion or the perplexity of the model, which is a posterior deviance measure (D. M. Blei et al., 2003). Various information criterion indices, such as the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002) the Akaike Information Criterion (AIC; Sakamoto et al., 1986), and the Bayesian Information Criterion (BIC; Raftery, 1995), have been used to select the number of topics. Mardones-Segovia et al. (2021) compared the accuracy of multiple model

selection methods using a simulation study and found that the Jensen-Shannon Divergence (JSD; Deveaud et al., 2014) and the sample adjusted BIC (SABIC; Sclove, 1987) metrics performed the best for selecting the correct number of topics under various educational measurement conditions. Additionally, it is important to note that the number of topics selected in educational measurement settings should also consider the interpretability and fairness of the results (e.g., Wheeler, Engelhard, et al., 2022a).

#### Application of LDA

Once the word probabilities have been estimated from the posterior distribution, the topic proportions for new, unseen documents are estimated through a Gibbs sampling algorithm where the posterior topic distributions are fixed. The topic assignments for the new documents are then estimated through a Gibbs sampling algorithm and the topic proportions are analytically derived from the topic assignments. Since the topics are fixed, the estimated topic proportions of the new documents are projected in the same semantic space as the original documents.

## 2.5 Methodology for Comparing Semantic Spaces

The previous section described, in detail, the LSA and LDA topic models. Both topic models described use the same observed data to estimate the latent semantic space of a corpus, however, the results of the models express the semantic spaces differently. Specifically, the LSA topic model uses latent topics to represent the semantic space and the LDA topic model uses model parameters to represent the semantic space. Therefore, the semantic spaces derived from these two models cannot be directly compared. This section introduces a methodology for comparing the estimated semantic spaces between the two topic models.

### 2.5.1 Comparing LSA and LDA Topic Models

LSA and LDA estimate and express the underlying latent semantic space of a corpus using different techniques. LSA expresses the semantic space as orthonormal topics (orthogonal and normalized), where each element of the topics can be any real number and each topic has a relative weight. On the other hand, LDA expresses the semantic space as proportions and probabilities using the Dirichlet distribution. In addition to expressing the semantic spaces differently, the number of topics often varies drastically between the two models (e.g., Wheeler, Wang, Tan, & Cohen, 2022).. LSA often uses a large number of topics whereas LSA uses a small number of topics.

Since LSA and LDA represent the latent semantic space differently, a straightforward comparison of the semantic structures is infeasible. Rather, the two models need to be scaled to one another. One way to do this is to compare each document to all other documents under each model and then compare those comparisons between the two models. A cosine similarity measure is a common metric used in topic models that compares the similarity between two documents with respect to the model's underlying latent semantic space (Li & Han, 2013). In this way, the cosine similarity measure can be used to compare the similarity between the documents under each model and then those comparisons can be compared using a cosine similarity measure.

### 2.5.2 Cosine Similarity

Cosine similarity measures the angle between two vectors in any high-dimensional space. The value of the cosine measure indicates how similar, or close, two vectors are in their high dimensional space. Suppose we want to compare two documents, $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ where $i, j \in \{1, ..., D\}$, from a corpus. Under the LSA topic model, the two documents can be expressed using the approximated document-by-word matrix, $\boldsymbol{X}'_{D\times V}$, as described in Equation

(**??**), where $\boldsymbol{X}'_{i,*}$ is document $\boldsymbol{w}_i$ projected in the semantic space obtained by LSA and $\boldsymbol{X}'_{j,*}$ is document $\boldsymbol{w}_j$ projected in the semantic space obtained by LSA. The cosine similarity between any two documents under the LSA topic model is given by

$$\begin{aligned} cos(\boldsymbol{X}'_{i,*}, \boldsymbol{X}'_{j,*}) &= \frac{\boldsymbol{X}'_{i,*} \times \boldsymbol{X}'_{j,*}}{||\boldsymbol{X}'_{i,*}|| \times ||\boldsymbol{X}'_{j,*}||} \\ &= \frac{\sum_{v=1}^{V}(x'_{i,v} \times x'_{j,v})}{\sqrt{\sum_{v=1}^{V} {x'_{i,v}}^2} \times \sqrt{\sum_{v=1}^{V} {x'_{j,v}}^2}}, \end{aligned} \tag{2.13}$$

where $x'_{i,v}$ is approximated use of word $v$ from the vocabulary in document $\boldsymbol{w}_i$; and $x'_{j,v}$ is approximated use of word $v$ from the vocabulary in document $\boldsymbol{w}_j$.

Under the LDA topic model, documents are expressed using their topic proportions, $\boldsymbol{\tau}_d$ $_d where d \in \{1, ..., D\}$. The cosine similarity between two documents, $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ where $i, j \in \{1, ..., D\}$, under the LDA topic model is given by

$$\begin{aligned} cos(\boldsymbol{\tau}_i, \boldsymbol{\tau}_j) &= \frac{\boldsymbol{\tau}_i \times \boldsymbol{\tau}_j}{||\boldsymbol{\tau}_i|| \times ||\boldsymbol{\tau}_j||} \\ &= \frac{\sum_{k=1}^{K}(\tau_{i,k} \times \tau_{j,k})}{\sqrt{\sum_{k=1}^{K} \tau_{i,k}^2} \times \sqrt{\sum_{k=1}^{K} \tau_{j,k}^2}}, \end{aligned} \tag{2.14}$$

where $\tau_{i,k}$ is the estimated proportion for the usage of topic $k$ in document $\boldsymbol{w}_i$; and $\tau_{j,k}$ is the estimated proportion for the usage of topic $k$ in document $\boldsymbol{w}_j$.

The cosine similarity measure is similar to a correlation measure and ranges from $(-1, 1)$. As two vectors (i.e., documents) become more similar, the cosine similarity approaches 1, where a cosine similarity of 1 means that the two documents are perfectly similar. As two vectors (i.e., documents) become more dissimilar, the cosine similarity approaches $-1$, where a cosine similarity of $-1$ means that the two documents are perfectly dissimilar.

**Cosine Similarity for the LSA Model**

Suppose the LSA model reduces the semantic space of a corpus to $K_{LSA}$ topics. The cosine similarity between every document in the semantic space obtained by LSA can be expressed as a $D \times D$ matrix, $\boldsymbol{S}_{LSA}$, where row, $i \in \{1, ..., D\}$, and column, $j \in \{1, ..., D\}$, are associated to documents $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$, respectively. The cosine similarity between every document is

$$\boldsymbol{S}_{LSA} = \begin{bmatrix} cos(\boldsymbol{X}'_{1,*}, \boldsymbol{X}'_{1,*}) & cos(\boldsymbol{X}'_{1,*}, \boldsymbol{X}'_{2,*}) & \dots & cos(\boldsymbol{X}'_{1,*}, \boldsymbol{X}'_{D,*}) \\ cos(\boldsymbol{X}'_{2,*}, \boldsymbol{X}'_{1,*}) & cos(\boldsymbol{X}'_{2,*}, \boldsymbol{X}'_{2,*}) & \dots & cos(\boldsymbol{X}'_{2,*}, \boldsymbol{X}'_{D,*}) \\ \vdots & \vdots & \ddots & \vdots \\ cos(\boldsymbol{X}'_{D,*}, \boldsymbol{X}'_{1,*}) & cos(\boldsymbol{X}'_{D,*}, \boldsymbol{X}'_{2,*}) & \dots & cos(\boldsymbol{X}'_{D,*}, \boldsymbol{X}'_{D,*}) \end{bmatrix}.$$

**Cosine Similarity for the LDA Model**

Suppose the LDA model reduces the semantic space of a corpus to $K_LDA$ topics. The cosine similarity between every document in the semantic space obtained by LSA can be expressed as a $D \times D$ matrix, $\boldsymbol{S}_{LDA}$, where row, $i \in \{1, ..., D\}$, and column, $j \in \{1, ..., D\}$, are associated to documents $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$, respectively. The cosine similarity between every document is

$$\boldsymbol{S}_{LDA} = \begin{bmatrix} cos(\boldsymbol{\tau}_1, \boldsymbol{\tau}_1) & cos(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2) & \dots & cos(\boldsymbol{\tau}_1, \boldsymbol{\tau}_D) \\ cos(\boldsymbol{\tau}_2, \boldsymbol{\tau}_1) & cos(\boldsymbol{\tau}_2, \boldsymbol{\tau}_2) & \dots & cos(\boldsymbol{\tau}_2, \boldsymbol{\tau}_D) \\ \vdots & \vdots & \ddots & \vdots \\ cos(\boldsymbol{\tau}_D, \boldsymbol{\tau}_1) & cos(\boldsymbol{\tau}_D, \boldsymbol{\tau}_2) & \dots & cos(\boldsymbol{\tau}_D, \boldsymbol{\tau}_D) \end{bmatrix}.$$

### 2.5.3 Semantic Space Similarity Metric

The $\boldsymbol{S}_{LSA}$ matrix expresses the similarity between every document using the semantic space obtained by LSA and the $\boldsymbol{S}_{LDA}$ matrix expresses the similarity between every document

using the semantic space obtained by LDA. Although the semantic spaces obtained from LSA and LDA differ in dimensionality and use a different scale, the cosine similarity matrices, $\boldsymbol{S}_{LSA}$ and $\boldsymbol{S}_{LDA}$, have the same dimensionality and scale. Moreover, the interpretation of the two matrices are similar, that is, both matrices contains information about the similarities between documents with respect to the semantic spaces obtained by both models. Therefore, to see if the semantic spaces between the two models capture similar information, another cosine similarity measure can be calculated between the same row in each matrix. That is, row $i$ in $\boldsymbol{S}_{LSA}$ and $\boldsymbol{S}_{LDA}$ are associated with document $\boldsymbol{w}_i$ and contains information about how similar document $\boldsymbol{w}_i$ is to all other documents based on the semantic space obtained by the LSA and LDA model, respectively. Calculating the cosine similarity between $\boldsymbol{S}_{LSA}$ and $\boldsymbol{S}_{LDA}$ gives a index for the similarity between the semantic spaces obtained by the LSA and LDA models:

$$S_i = cos(\boldsymbol{S}_{LSA\{i,*\}}, \boldsymbol{S}_{LDA\{i,*\}}), \tag{2.15}$$

where $S_i$ represents the cosine similarity between document $\boldsymbol{w}_i$ in the semantic spaces obtained by LSA and LDA. Taking the average cosine similarity across all documents, $\bar{S} = \frac{1}{D}\sum_{i=1}^{D} S_i$, provides a proxy for the similarity between the semantic spaces obtained by LSA and LDA.

## 2.6 Simulation Study

In this section, we use a simulation study to investigate and compare the semantic spaces for LSA and LDA using the methodology described in the previous section.

### 2.6.1 Simulation Design and Conditions

The simulation design considered realistic scenarios found in previous topic model studies of constructed-response items (e.g., Choi et al., 2017; S. Kim et al., 2017; Wheeler, Engelhard,

et al., 2022a; Xiong et al., 2020). A total of 5 factors were manipulated: number of topics in LDA (three levels: 3, 4, and 5 topics), number of topics in LSA (three levels: 50, 100, and 150 topics), number of essays (three levels: 250, 750, and 1250 documents), vocabulary size (two levels: 750 and 1250 words), and the average length of the essays (two levels: 50 and 150 words). All simulation factors were crossed for a total of 108 conditions and each condition was ran on 25 replications.

### 2.6.2 Data Generation

The data for the simulation study were generated using the assumed LDA generative process. First, each topic $k \in \{1, ..., K\}$ was generated from a Dirichlet distribution with a symmetric hyperparameter set to 0.25, such that $\boldsymbol{\beta}_k \sim Dirichlet(0.25) \ \forall \ k \in \{1, ..., K\}$. The topic proportions for each document $d \in \{1, ..., D\}$ were generated from a Dirichlet distribution with a symmetric hyperparameter set to 0.25, such that $\boldsymbol{\tau}_d \sim Dirichlet(0.25) \ \forall \ d \in \{1, ..., D\}$. Then, each document is generated by first generating the length from a Poisson distribution with the parameter set to the average length simulation condition (50 or 150), such that $N_d \sim Poisson(\lambda)$ where $\lambda = 50, 150$. Then, the topic assignments for each word in each document are generated from a multinomial distribution with the parameter set to the topic proportions of the document, such that $\boldsymbol{t}_d \sim Multinomial(\boldsymbol{\tau}_d)$. Finally, the words for each document are generated from a multinomial distribution using the topics and topic assignments of the document, such that $\boldsymbol{w}_d \sim Multinomial(\boldsymbol{\beta}_{1:K}|\boldsymbol{t}_d)$.

### 2.6.3 Parameter Recovery

Since the data were generated using the generative process and probabilistic model of LDA, the recovery of model parameters were checked to ensure that the estimated models could be compared. For each condition, a cosine similarity measure was used to compare the estimated topics and topic proportions with the known topics and topic proportions that

were used to generate the data. A cosine similarity measure greater than 0.9 indicates the parameters of the model were successfully recovered and the results of the study can be interpreted; a cosine similarity measure between 0.8 and 0.9 indicates that the parameters were partially recovered and the results of the study should be interpreted with caution; and a cosine similarity measure less than 0.7 indicates that the parameters were not recovered and the results of the study should not be interpreted (Wheeler, Xiong, et al., 2022). Table 2.1 contains the results for the average cosine similarity between the estimated and known parameters of the model for each condition across the 25 replications.

The parameters of the model were recovered under most of the simulation conditions, except for the conditions where the sample size was 250, the vocabulary length was 1500, and the average essay length was 50, which had cosine similarities below 0.8. One possible explanation for the lower similarities under this condition is the number of parameters being estimated and the lack of data, which is consistent with other LDA simulations with similar conditions (e.g., Mardones-Segovia et al., 2022; Wheeler, Cohen, et al., 2021; Wheeler, Xiong, et al., 2022). It is also important to note that the recovery of parameters was not necessarily dependent on a single condition, but rather the combination of two main properties of the corpus: sample size and average essay length. That is, if the average essay length is large then fewer documents are needed, but if the average essay length is small then more documents are needed. This is a consistent finding within the simulation literature for topic models (Wild, 2016, see chapter 7).

### 2.6.4 Simulation Results

For each simulation condition and replication, data are generated through the described generative process, and the LSA and LDA topic models are fit to the generated data. The semantic spaces for both models were compared using the method described in Section 2.5. Table 2.2 summarizes the comparison between the semantic spaces obtained by LSA and the

Table 2.1: Parameter recovery of the LDA model for each condition

| Number of Topics | Sample Size | Vocab Length | Essay Length | Avg. Cosine Similarity of Word Probabilities | Avg. Cosine Similarity of Topic Proportions |
|---|---|---|---|---|---|
| 3 | 250 | 750 | 50 | 0.878 (0.028) | 0.871 (0.029) |
| | | | 150 | 0.962 (0.010) | 0.961 (0.010) |
| | | 1500 | 50 | 0.765 (0.036) | 0.736 (0.041) |
| | | | 150 | 0.945 (0.011) | 0.939 (0.011) |
| | 750 | 750 | 50 | 0.910 (0.021) | 0.901 (0.022) |
| | | | 150 | 0.961 (0.012) | 0.959 (0.012) |
| | | 1500 | 50 | 0.864 (0.026) | 0.837 (0.030) |
| | | | 150 | 0.962 (0.009) | 0.957 (0.009) |
| | 1250 | 750 | 50 | 0.917 (0.020) | 0.906 (0.021) |
| | | | 150 | 0.963 (0.014) | 0.962 (0.014) |
| | | 1500 | 50 | 0.879 (0.022) | 0.852 (0.025) |
| | | | 150 | 0.960 (0.010) | 0.956 (0.010) |
| 4 | 250 | 750 | 50 | 0.857 (0.028) | 0.847 (0.029) |
| | | | 150 | 0.941 (0.015) | 0.939 (0.015) |
| | | 1500 | 50 | 0.739 (0.040) | 0.704 (0.045) |
| | | | 150 | 0.924 (0.015) | 0.915 (0.016) |
| | 750 | 750 | 50 | 0.880 (0.025) | 0.867 (0.025) |
| | | | 150 | 0.945 (0.016) | 0.942 (0.016) |
| | | 1500 | 50 | 0.842 (0.028) | 0.809 (0.031) |
| | | | 150 | 0.942 (0.013) | 0.935 (0.013) |
| | 1250 | 750 | 50 | 0.897 (0.021) | 0.884 (0.022) |
| | | | 150 | 0.948 (0.015) | 0.946 (0.015) |
| | | 1500 | 50 | 0.856 (0.026) | 0.826 (0.030) |
| | | | 150 | 0.942 (0.015) | 0.938 (0.014) |
| 5 | 250 | 750 | 50 | 0.827 (0.029) | 0.815 (0.031) |
| | | | 150 | 0.932 (0.019) | 0.929 (0.019) |
| | | 1500 | 50 | 0.718 (0.039) | 0.680 (0.042) |
| | | | 150 | 0.900 (0.016) | 0.890 (0.016) |
| | 750 | 750 | 50 | 0.867 (0.026) | 0.851 (0.027) |
| | | | 150 | 0.937 (0.017) | 0.935 (0.017) |
| | | 1500 | 50 | 0.812 (0.030) | 0.775 (0.033) |
| | | | 150 | 0.924 (0.018) | 0.917 (0.018) |
| | 1250 | 750 | 50 | 0.879 (0.026) | 0.865 (0.027) |
| | | | 150 | 0.932 (0.019) | 0.930 (0.019) |
| | | 1500 | 50 | 0.827 (0.028) | 0.792 (0.031) |
| | | | 150 | 0.928 (0.018) | 0.922 (0.018) |

known generated LDA model. The results suggest that the semantic spaces for both models are highly similar under most conditions. This infers that the semantic spaces obtained by the LSA and LDA models capture similar and consistent information about the corpus even though the semantic spaces are obtained differently and expressed differently.

Table 2.2 identifies three main findings about the similarities between the semantic spaces obtained by LSA and LDA. First, there is a positive correlation between the sample size and the similarity between semantic spaces, since an increase in sample size means there is more data and more information regarding the relationship between documents and words (i.e., more information about the latent semantic space). That is, the semantic spaces become clearer as the amount of data increases which leads to high similarities between the semantic spaces obtained by the models. Second, there is a negative correlation between vocabulary size and the similarity between semantic spaces. Lastly, there is a positive correlation between the average length of the documents and the similarity between the semantic spaces. Wild (2016) found similar relationships between these properties of the corpus and the semantic space obtained by LSA.

## 2.7 Empirical Study

The previous section investigated the semantic spaces obtained from LSA and LDA using a simulation study. This section uses the same methodology to compare the semantic spaces obtained by LSA and LDA using data collected from a sample of responses to a CR item on a high school English Language Arts (ELA) assessment.

### 2.7.1 Methodology

#### Data Source and Description

A sample of $3,815$ written responses to a CR item from a ninth-grade ELA formative writing assessment was used for this study. The item prompted the students to write an informational essay about the effects of social media and social networking. Each student's response is constituted as a document, thus $D = 3815$. The average document length was 320.43 words. The total number of words that appeared in the corpus was $1,187,835$ and the length of the vocabulary was $V = 4318$.

#### Preprocessing Steps and Data Cleaning

There are several data-cleaning tasks before fitting a topic model to empirical data (Boyd-Graber et al., 2014). First, all punctuation is removed from the documents and all words are converted to lowercase, which ensures that a lowercase and uppercase version of the same word is not counted as two unique words. This step is often referred to as *tokenization* in the natural language processing literature. Next, stop words are removed from the documents. Stop words are high-frequency words that carry little information about the semantic space (e.g., *the*, *what*, *are*, *is*). Lastly, words are stemmed which retrieves the base form of a word (e.g., *run* is the stemmed version of *runs*, *running*, *ran*). This step is often referred to as *lemmatization* in the natural language processing literature. These data-cleaning steps are crucial as they increase the clarity of the semantic spaces and improve the interpretation of the results (Schofield et al., 2017).

After data-cleaning steps, the updated number of documents changed to $D = 3698$ (117 documents were blank after removing stop words); the updated average length of a document was 186.50 words; the updated total number of words in the corpus was $689,667$ and the updated length of the vocabulary was $V = 3607$. Table 2.3 displays the change in these

properties of the corpus throughout each data-cleaning step. It is important to note that the changes in the corpus during the data-cleaning steps are sample dependent.

#### Model Estimation

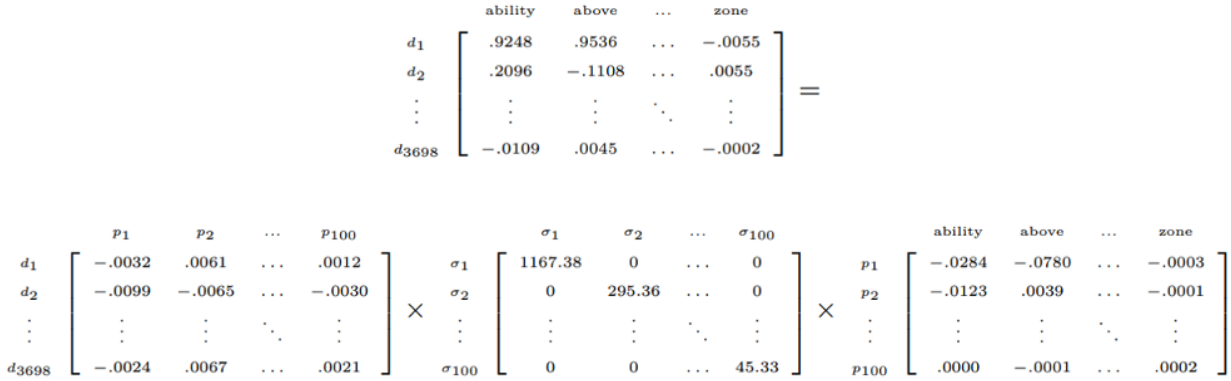
After the data-cleaning steps, each document $d = 1, ..., D = 3698$ was mathematically represented by $\boldsymbol{w}_d$, following Section 2.1. The LSA and LDA models are applied to the corpus following the steps outlined in Sections 2.2 and 2.3. Both topic models were estimated using the R programming language. The *lsa* package (Wild, 2007) was used for LSA and the *topicmodels* package (Grün & Hornik, 2011) was used for LDA.

### 2.7.2 Empirical Results

The LSA model reduced the semantic space to 100 topics. Figure 2.4 displays the results of the LSA reduction. The approximated document-by-word (top matrix) shows the projection of the original document-by-word matrix on the reduced semantic space of 100 topics. The reduced document matrix (bottom left matrix) indicates how the documents load on the 100 topics. The reduced word matrix (bottom right matrix) indicates how the words load on each topic. The weights for the 100 topics are shown in the reduced topic matrix (middle diagonal matrix), which shows that the first topic has a weight of $\sigma_1 = 1167.38$ and the last topic has a weight of $\sigma_{100} = 45.33$.

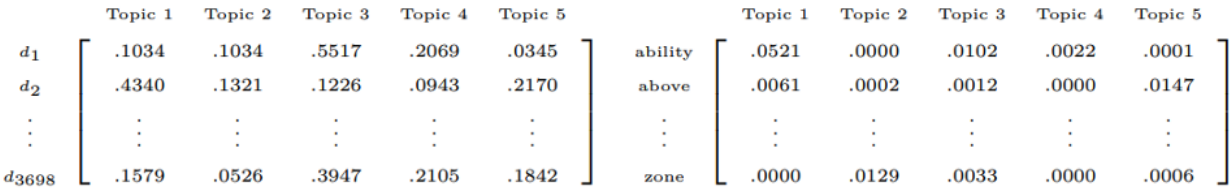
The LDA model reduced the semantic space to 5 topics. Figure 2.5 displays the results of the LDA estimation, specifically the topics and topic proportion parameters. The topic proportions (left matrix) show the usage of each topic for each document. For example, the topic proportion for the first document (first row) is $\boldsymbol{\tau}_1 = (.1034, .1034, .5517, .2069, .0345)$ which means the first document used Topic 1 10%, Topic 2 10%, Topic 3 55%, Topic 4 21%, and Topic 5 4%. The topics (right matrix) show the probability of each word in the

Figure 2.4: LSA numerical results for a sample of 3698 documents and 100 topics



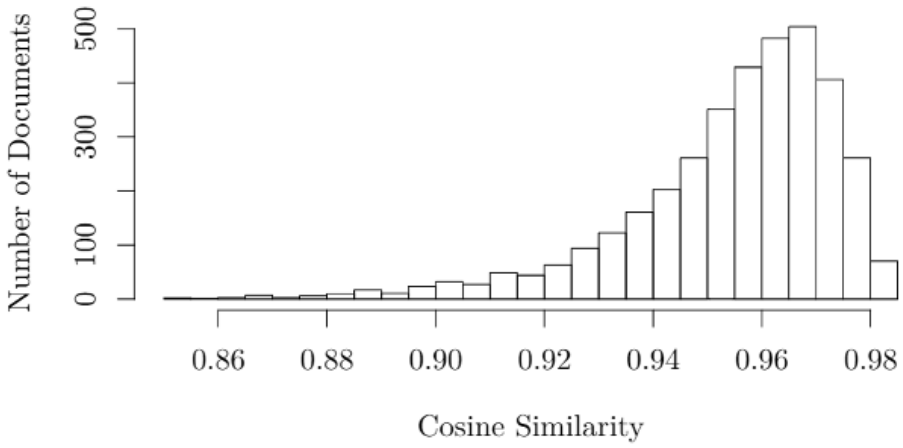

vocabulary occurring under that topic. For example, given Topic 1, the word "ability" has 5% chance of appearing.

Figure 2.5: LDA numerical results for a sample of 3698 documents and 5 topics




The semantic spaces obtained from LSA and LDA were compared using the methodology outlined in Section 2.5. Figure 2.6 shows a histogram of the cosine similarities for each document under the topic models. Most of the documents ($n = 3405$; 92.1%) had cosine similarities greater than 0.90 and only a few essays ($n = 4$; 0.1%) had cosine similarities less than 0.50. The average cosine similarity between the LSA and LDA models was $\bar{S} = 0.949$ (0.021). This suggests that the semantic spaces obtained by LSA and LDA are highly similar and capture similar information.

Figure 2.6: Histogram of the cosine similarities for each document between LSA and LDA




## 2.8 Discussion

Topic models provide potentially useful means of analyzing the underlying semantic space of texts. LSA has been used successfully as a basis for automatic essay scoring, as it has been shown to have high consistency between human and machine scores. LDA has been used successfully in a wide variety of fields as a method for detecting the latent thematic structure of large corpora of text. LDA has only recently been applied for use in analyzing the text of test takers' answers to constructed responses (e.g., Choi et al., 2017; S. Kim et al., 2017; Wheeler, Raczynski, et al., 2022). Table 2.4 provides an overview of the key distinctions between the LSA and LDA models described throughout this paper.

This paper reviewed the mathematical models and estimation procedures for LSA and LDA, and described how these models differ theoretically. LSA uses a document-by-word matrix to represent the corpus whereas LDA uses an assumed statistical model. Within the context of educational measurement, LSA typically reduces the semantic space to a large number of topics whereas LDA typically reduces the semantic space to a small number of topics. Additionally, due to the assumed statistical model and the scale of the parameters,

the results of LDA are typically easier to interpret and explain than the results of LSA. Based on the current educational measurement literature, LSA has traditionally been used as the basis for automated scoring algorithms whereas LDA is typically used to extract additional information about students' responses beyond the score.

There are a limited number of studies that have investigated the theoretical similarities between these two topic models. Therefore, this study investigated how these two models compare in their mathematical setup and how they represent the semantic space of a corpus through a simulation and empirical example. Since the two topic models express the semantic spaces using different mathematical representations, we introduced a methodology that calculates the similarity (or correlation) between all documents under each model and then calculates a similarity measure of the sets of similarity values under each model, which determines how similar the semantic spaces are relative to the textual data sample. In other words, we use LSA to rank students and LDA to rank students and then we compare the rankings between the two models. This comparison method provides a measure of the internal consistency between the LSA and LDA semantic spaces.

The simulation study was conducted to generalize the comparison between LSA and LDA using common educational measurement conditions. The results of this comparison suggest that the estimated semantic spaces derived from LSA and LDA are highly similar, and that the two models infer relatively similar semantic spaces but express them differently. Additionally, the simulation suggested that the similarity between the semantic spaces is not dependent on the words or context, but rather the number of documents and words observed.

Although LSA and LDA estimate relatively similar semantic spaces, they are used for different purposes in educational measurement. LSA is typically used for automated essay scoring and extracting features for other machine learning models. LDA has typically been used to assess students' thinking and reasoning when responding to CR items and to gain additional information beyond the scores. LDA performs poorly for automated essay scoring

and typically leads to larger errors and lower human-machine score agreement. This is counterintuitive since the semantic spaces were found to be highly similar, however, the mathematical formulation of these models can shed light on this.

The features extracted using LSA are orthonormal topics and have no estimation error because they are obtained directly through the decompositions of the document-by-word matrix, which is beneficial for automated essay scoring because it eliminates a potential source of measurement error. On the other hand, the features extracted using LDA are estimated through an assumed statistical model, thus the parameter estimates naturally contain measurement error. Additionally, the assumed topics and topic proportions are assumed to follow a Dirichlet distribution, which is naturally ipsative. That is, the topics and topic proportions must sum to 1, which means that topics cannot have high probabilities on a large number of words and topic proportions cannot have high probabilities on a large number of topics. This is a restrictive quality that causes a dependency within topics and topic proportions. Therefore, when the features from LDA are used to replicate scores for essays, the automated scoring algorithms contain an additional source of error and the features may not be linearly related to the scores.

The results of the study also suggest that these two methods can be augmented together to provide a more complete analysis of students since they analyze the same set of data. That is, the LSA method is used in many automated essay scoring algorithms due to its high reliability and consistency with human raters while the LDA method is used to gain more granular information about the students' thinking and reasoning when responding to CR items, that may not be captured by the rubric scores. Although these two models are traditionally used for separate purposes, the semantic spaces obtained by both models are highly similar. This suggests that these two models can be augmented to increase the quality of information obtained when analyzing CR items.

Table 2.2: Simulation results for comparing the semantic structure between LSA and LDA for various constructed-response item scenarios

| Number of Topics | Sample Size | Vocab Length | Essay Length | Cosine Similarity of LSA to True LDA | | |
|---|---|---|---|---|---|---|
| | | | | k = 50 | k = 100 | k = 150 |
| 3 | 250 | 750 | 50 | 0.895 (0.025) | 0.878 (0.028) | 0.873 (0.029) |
| | | | 150 | 0.965 (0.010) | 0.962 (0.010) | 0.961 (0.010) |
| | | 1500 | 50 | 0.813 (0.032) | 0.765 (0.036) | 0.744 (0.039) |
| | | | 150 | 0.954 (0.010) | 0.945 (0.011) | 0.941 (0.011) |
| | 750 | 750 | 50 | 0.923 (0.020) | 0.910 (0.021) | 0.904 (0.022) |
| | | | 150 | 0.963 (0.013) | 0.961 (0.012) | 0.960 (0.012) |
| | | 1500 | 50 | 0.895 (0.024) | 0.864 (0.026) | 0.847 (0.028) |
| | | | 150 | 0.966 (0.009) | 0.962 (0.009) | 0.959 (0.009) |
| | 1250 | 750 | 50 | 0.930 (0.019) | 0.917 (0.020) | 0.910 (0.020) |
| | | | 150 | 0.965 (0.014) | 0.963 (0.014) | 0.963 (0.014) |
| | | 1500 | 50 | 0.907 (0.020) | 0.879 (0.022) | 0.863 (0.024) |
| | | | 150 | 0.964 (0.011) | 0.960 (0.010) | 0.958 (0.010) |
| 4 | 250 | 750 | 50 | 0.878 (0.026) | 0.857 (0.028) | 0.850 (0.029) |
| | | | 150 | 0.944 (0.015) | 0.941 (0.015) | 0.939 (0.015) |
| | | 1500 | 50 | 0.795 (0.037) | 0.739 (0.040) | 0.714 (0.044) |
| | | | 150 | 0.937 (0.015) | 0.924 (0.015) | 0.918 (0.016) |
| | 750 | 750 | 50 | 0.897 (0.025) | 0.880 (0.025) | 0.872 (0.025) |
| | | | 150 | 0.947 (0.016) | 0.945 (0.016) | 0.943 (0.016) |
| | | 1500 | 50 | 0.880 (0.026) | 0.842 (0.028) | 0.822 (0.029) |
| | | | 150 | 0.948 (0.013) | 0.942 (0.013) | 0.938 (0.013) |
| | 1250 | 750 | 50 | 0.913 (0.021) | 0.897 (0.021) | 0.889 (0.022) |
| | | | 150 | 0.951 (0.015) | 0.948 (0.015) | 0.947 (0.015) |
| | | 1500 | 50 | 0.887 (0.025) | 0.856 (0.026) | 0.838 (0.029) |
| | | | 150 | 0.947 (0.015) | 0.942 (0.015) | 0.940 (0.014) |
| 5 | 250 | 750 | 50 | 0.852 (0.028) | 0.827 (0.029) | 0.818 (0.030) |
| | | | 150 | 0.936 (0.019) | 0.932 (0.019) | 0.930 (0.019) |
| | | 1500 | 50 | 0.781 (0.039) | 0.718 (0.039) | 0.692 (0.041) |
| | | | 150 | 0.914 (0.017) | 0.900 (0.016) | 0.893 (0.016) |
| | 750 | 750 | 50 | 0.888 (0.026) | 0.867 (0.026) | 0.856 (0.026) |
| | | | 150 | 0.940 (0.018) | 0.937 (0.017) | 0.935 (0.017) |
| | | 1500 | 50 | 0.852 (0.029) | 0.812 (0.030) | 0.789 (0.031) |
| | | | 150 | 0.931 (0.019) | 0.924 (0.018) | 0.920 (0.018) |
| | 1250 | 750 | 50 | 0.896 (0.026) | 0.879 (0.026) | 0.870 (0.026) |
| | | | 150 | 0.935 (0.020) | 0.932 (0.019) | 0.931 (0.019) |
| | | 1500 | 50 | 0.863 (0.028) | 0.827 (0.028) | 0.806 (0.029) |
| | | | 150 | 0.933 (0.019) | 0.928 (0.018) | 0.924 (0.018) |

Table 2.3: Corpus descriptive statistics after each data cleaning step

| | Before Data Cleaning | Removing Stop Words | Lemmatisation |
|---|---|---|---|
| **Documents** | 3,815 | 3,698<br>(~ 3% removed) | 3,698<br>(0% removed) |
| **Average Length** | 320.43 | 219.21<br>(~32% removed) | 186.50<br>(~15% remove) |
| **Total Observed Words** | 1,187,835 | 810,698<br>(~32% removed) | 689,667<br>(~15% removed) |
| **Vocabulary Size** | 4,318 | 4,202<br>(~ 3% removed) | 3,607<br>(~14% removed) |

Table 2.4: Overview of key features from Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) models

| | Latent Semantic Analysis (LSA) | Latent Dirichlet Allocation (LDA) |
|---|---|---|
| **Mathematical Foundation** | Linear Algebra,<br>Matrix Algebra | Assumed Probability Model<br>Assumed Generative Process |
| **Corpus Representation** | Document-by-Word Matrix | Probability Distributions<br>(Dirichlet) |
| **Estimation Procedures** | Singular Value Decomposition<br>Orthonormal Eigenvectors | Variational Bayes,<br>Gibbs Sampling,<br>EM algorithm |
| **Topic Interpretation** | Eigenvectors,<br>Eigenvalues | Probabilities |
| **Number of Topics in Educational Measurement** | 50-150 Topics | 2-10 Topics |
| **Educational Measurement Purposes** | Essay Scoring Algorithms,<br>Plagiarism Detection,<br>Process data extraction method | Assessing rater accuracy,<br>Students' thinking and reasoning,<br>Determining fairness of rubrics,<br>Quick qualitative analyses |

# Chapter 3

# Exploring Rater Accuracy Using Unfolding Models Combined with Topic Models: Incorporating Supervised Latent Dirichlet Allocation[2]

[2] Wheeler, J. M., Engelhard, G., & Wang, J. Submitted to *Measurement: Interdisciplinary Research and Perspectives*, 2/2/2022

## 3.1 Preface

The study presented in this chapter is currently under revision in Wheeler, Engelhard, et al. (2022b). The study was directed under the supervision of Drs. George Englehard and Jue Wang. My role in this study was to develop the R code to fit the topic models, analyze the results from the topic models in conjunction with the results from the unfolding model, and draft the initial manuscript. This paper demonstrates the utility of topic models by showing how they can be used to provide a substantive interpretation of latent scales estimated by item response models.

## 3.2 Abstract

Objectively scoring extended-response items on educational assessments has long been a challenge because of the use of human raters. Even well-trained raters using a rubric can inaccurately assess essays. There might be different sources causing this inaccuracy in scoring. Unfolding models measure rater's scoring accuracy that captures the discrepancy between criterion and operational ratings by placing essays on an unfolding continuum with an ideal-point location. The probability of accurately scoring an essay decreases as a rater's location moves away from the essay's ideal point on the unfolding scale. Essay unfolding locations can indicate how difficult it is for raters to score an essay accurately. This study aims to explore a substantive interpretation of the unfolding scale based on a supervised Latent Dirichlet Allocation (sLDA) model. We investigate the relationship between latent topics extracted using sLDA and unfolding locations with a sample of essays ($n = 100$) obtained from an integrated writing assessment. Results show that (a) three latent topics moderately explain ($r^2 = 0.561$) essay locations defined by the unfolding scale and (b) failing to use and/or cite the source articles led to essays that are difficult-to-score accurately. This information can

be used to improve operational scoring practices by identifying potential sources of rater inaccuracy.

## 3.3 Introduction

Rater-mediated assessments contain responses that require rater scoring, such as extended-response items, performance tasks, and portfolios. These assessments often allow students to better demonstrate their competencies, and are frequently used to assess high-order thinking skills. Even if raters are thoroughly trained and follow a rubric, these rater-mediated assessments are still susceptible to potential biases, ranging from social positioning biases, such as gender, race, and class, to systematic biases, such as severity or leniency, and central tendency (Engelhard Jr, 1994; Read et al., 2005; Saal et al., 1980). Nonetheless, these potential biases add a source of variance to the ratings that are irrelevant to the construct being measured and infringe on the validity and fairness of the ratings (Eckes, 2005; Rezaei & Lovorn, 2010).

Detecting rater biases can be achieved through various statistical methods (Aubin et al., 2018; Engelhard, 2012; Myford & Wolfe, 2003). These methods rely on a model-data fit approach that identifies raters that exhibit aberrant response patterns. Although quantitative approaches are able to identify biases in rater-mediated assessments, further analyses are necessary for an evaluation of the causes for raters to elicit biases toward particular essays. A method that is able to identify and evaluate sources of inaccuracy would help policymakers prescribe various interventions to monitor and mitigate rater biases.

Conceptualizing rater judgments and the rating process, however, is a challenge. Previous research developed different models for examining rater effects, including latent trait modeling (Engelhard Jr, 1994; Wolfe & McVay, 2012), hierarchical rater model (Patz et al., 2002), rater Bundle Model (Wilson & Hoskens, 2001), and generalized rater model (W.-C. Wang et al., 2014). Recent studies have tried to understand this process through an unfolding model (J. Wang & Engelhard Jr, 2019a, 2019b; J. Wang et al., 2016). These studies examined individual

differences among raters in scoring the essays. For instance, raters may show different levels of accuracy toward scoring the same essay. In other words, an essay may appear to be difficult-to-score across different raters. Therefore, instead of assuming that raters score the essays in a consistent manner, unfolding models examine individual differences among raters in scoring essays and help us understand to what degree raters assess essays differently. When the focus is placed on the evaluation of rater accuracy, unfolding models define the essays based on their difficulty-to-score for individual raters.

A challenge when using an unfolding model to understand rater judgments is that the substantive interpretation of the unfolding scale is not easily defined. J. Wang and Engelhard Jr (2019b) used essay feature indices from Coh-Metrix to explore the meaning of an unfolding scale and found promising results. A quantitative analysis of the content of the essays through the use of topic models, which are a set of statistical models used to analyze textual data, could provide additional interpretations about the meaning of locations along the unfolding scale. The goal of topic models is to estimate latent topics found in a collection of essays. Latent topics are similar to principle components and represent clusters of words that have similar context across a collection of essays. Latent topics are used to infer the relationship between essays and words. A recent study showed that the results from topic models provide additional information beyond the score and provide insight into the writing process of students (Cardozo-Gaibisso et al., 2019). Topic models have also been used to analyze the effects of instructional writing interventions by identifying latent topics associated with different language use in students' essays (Duong et al., 2019; S. Kim et al., 2017). Therefore, topic models enable a robust analysis of textual data that may influence the scores assigned by raters to essays.

### 3.3.1 Purpose of Study

In this study, we used an unfolding model combined with a supervised topic model to evaluate essays written by students to an extended-response item in an integrated writing assessment. Integrated writing assessments require students to read source passages, and to utilize information from the passages in answering an essay prompt. Integrated writing assessments are becoming more common in statewide assessments. The unfolding model is used to define difficult-to-score essays, and the supervised topic model is used to estimate the latent topics from the content of the essays. The study investigates two questions: (1) what are the latent topics used across all essays? and (2) how do the latent topics relate to the unfolding locations of each essay? This information can be used to identify sources of inaccuracy based on the latent topics and ultimately improve rater training and the quality of ratings.

## 3.4 Methodology

### 3.4.1 Accuracy Ratings

The accuracy ratings directly reflect the distance between the criterion and observed ratings (Engelhard Jr, 1997; Engelhard Jr, 2013). They can be calculated based on the equation below:

$$A_{i,j} = max_{i=1,\ldots,I;j=1,\ldots,J}\{|O_{i,j} - C_i|\} - |O_{i,j} - C_i|, \tag{3.1}$$

where $J$ is the number of raters and $I$ refers to the number of student essays; $A_{ij}$ is the accuracy rating of rater $j$ on essay $i$; $O_{ij}$ is the raw rating of Rater $j$ on Essay $i$; $C_i$ is the criterion rating on Essay $i$. Criterion ratings may be defined by a panel of expert raters or the average of observed ratings. In this study, criterion ratings are given by expert raters.

Modeling this distance using the accuracy ratings allows us to create a latent continuum of rating accuracy (Engelhard Jr, 1996).

### 3.4.2 Unfolding Model

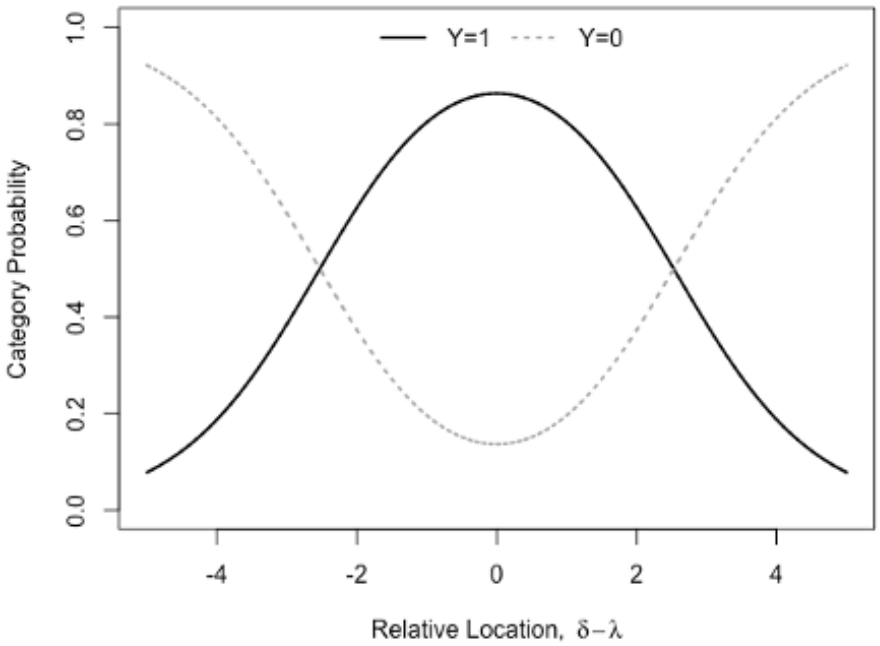
In this study, accuracy ratings are modeled using the hyperbolic cosine model (HCM; Andrich & Luo, 1993) to define the hyperbolic cosine accuracy model (HCAM; J. Wang et al., 2016) for creating an accuracy continuum. The HCAM is defined as

$$P(X_{i,j} = k) = \frac{[cosh(\delta_i - \lambda_j)]^{m-k} \prod_{l=1}^{k} cosh(\rho_{i,l})}{\sum_{k=0}^{m} [cosh(\delta_i - \lambda_j)]^{m-k} \prod_{l=1}^{k} cosh(\rho_{i,l})}, \tag{3.2}$$

where $m$ is the number of categories of accuracy ratings; $X_{i,j}$ is the observed accuracy rating received by rater $j$ on essay $i$; $\delta_i$ is the difficulty of essay $i$ to score accurately; $\lambda_j$ is the accuracy location of rater $j$; $\rho_{i,l}$ is the essay threshold parameter, which are constrained to be equally distanced across accuracy rating categories; and when $k = 0$ then $\prod_{l=1}^{k} cosh(\rho_{i,l}) \equiv 1$.

The HCAM explores the difficulty-to-score of essays for individual raters. In other words, raters with different accuracy locations on the unfolding scale may score different subsets of essays accurately. The HCAM achieves this by modelling the relative distance between the essays and raters based on a proximity principle. That said, essays and raters are located along a common unfolding scale, and raters locate closer to the essays that they score more accurately. Figure 3.1 shows the probability function curves for an essay based on HCAM. The relative location refers to the relative difference between a rater's location and an essay's location. Raters who are located closer to an essay have greater likelihood to score this essay accurately. With the increase in the relative locations, the probability of accurate scoring ($X = 1$) decreases and the probability of inaccurate scoring ($X = 0$) increases. In this study, the essay locations on the unfolding continuum are used for exploring the substantive meaning of the scale with the use of a supervised Latent Dirichlet Allocation model.

Figure 3.1: Probability function curves for a Hyperbolic Cosine Accuracy Model




*Note.* $Y = 1$ indicates that there is an agreement between operational ratings and expert ratings. $Y = 0$ indicates that there is not an agreement between operational ratings and expert ratings.

### 3.4.3 Supervised Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA; D. M. Blei et al., 2003) is a probabilistic mixture model that uses a hierarchical Bayesian design to estimate latent topics within a collection of texts. The hierarchical design works as follows: a collection of text forms the corpus, the corpus consists of a predetermined number of latent topics which are a weighted mixture over a predetermined vocabulary, and each text within the corpus consists of topic proportions which are a weighted mixture over the number of latent topics (D. Blei et al., 2010). For this study, the corpus is formed by a set of essays. The goal of LDA is to estimate the latent

topics used throughout the essays and the topic proportions of each essay. The latent topics are clusters of words that are related and represent common components seen across the collection of essays. The topic proportions sum to 1 and represent the usage of each topic. LDA estimates the latent topics and topic proportions through its posterior distribution shown in Equation (3.3) using a collapsed Gibbs sampling algorithm:

$$P(\boldsymbol{\tau}, \boldsymbol{\beta}, \boldsymbol{t}|w, \eta, \nu) \propto \prod_{n=1}^{N_d} P(t_{d,n}|\boldsymbol{\tau}_d) P(w_{d,n}|t_{d,n}, \boldsymbol{\beta}_{k=t_{d,n}}) \times \prod_{d=1}^{D} P(\boldsymbol{\tau}_d|\eta) \times \prod_{k=1}^{K} P(\boldsymbol{\beta}_k|\nu) \quad (3.3)$$

where $N_d$ represents the number of words in essay $d \in \{1, ..., D\}$; $\boldsymbol{\tau}_d$ represents topic proportions for each essay $d \in \{1, ..., D\}$; $\boldsymbol{\beta}_k$ represents topic $k \in \{1, ..., K\}$; $t_{d,n}$ represents the topic assignment for each word $n \in \{1, ..., N_d\}$ in each essay $d \in \{1, ..., D\}$; $w_{d,n}$ represents each of the observed words $n \in \{1, ..., N_d\}$ in each essay $d \in \{1, ..., D\}$; $\eta$ and $\nu$ are the prior hyperparameters of the model and represent the concentration of topic proportions in each essay and words within topics, respectively. A small $\eta$ indicates essays consist of mainly one topic and a large $\eta$ indicates essays consist of all topics uniformly. A small $\nu$ indicates that topics are associated with few words and a large $\nu$ indicates that topics are associated with all words uniformly (D. M. Blei & Lafferty, 2006). Since topics are typically unknown beforehand, noninformative prior hyperparameters are often chosen, that is, $\eta = 1$ and $\nu = 1$.

A restriction of LDA is that the estimation of the topics and topic proportions are unsupervised, that is, there are no additional variables that can help drive the inference. However, if we have a dependent variable, such as a score or unfolding location, a supervised latent Dirichlet allocation model (sLDA; Mcauliffe & Blei, 2007) can be estimated. The sLDA model uses the dependent variable, that is essay unfolding locations, to help drive inference of the topic assignments, $t$, from Equation (3.3). The dependent variable is related to the topic assignments through the following response distribution.

$$y_d|\boldsymbol{t}_d, \mu, \sigma^2 \sim Normal(\mu\bar{t}_d, \sigma^2), \tag{3.4}$$

where $y_d$ represents the dependent variable for essay $d \in \{1, ..., D\}$ and is assumed to follow a normal distribution with prior means $\mu$ and prior variance $\sigma^2$; $\bar{t}_d$ represents the relative use of each topic in each essay and is defined by $\bar{t}_d = \frac{1}{N_d}\sum_{n=1}^{N_d} t_{d,n}$.

The sLDA models the relationship between the topic proportions of each essay and the unfolding measures through a regression model. Results can indicate the degree to which each topic is associated with the essay unfolding measures.

### 3.4.4 Data Source

The essays were written responses to a seventh grade informational extended-response item from an English Language Arts (ELA) formative writing assessment. The extended-response item provided two passages, Passage A and Passage B, and asked the students to write an informative essay that answers the prompt by citing two examples from each passage as evidence. Please see the Appendix for Passages A and B.

Prior to this study, each essay was given an operational score by well-trained operational raters using a rubric-based scoring system. A sample of 100 essays were selected and given a criterion score by a group of expert raters. The distance between the operational scores and criterion scores were used to calculate the accuracy rating of each essay based on Equation (3.1), and the dichotomous accuracy ratings were analyzed with the HCAM (1 = accurate, 0 = inaccurate).

The sample of 100 essays were then used to fit a sLDA model using the lda R package (Chang & Chang, 2010). Prior to fitting the sLDA model, the essays were put through a data pipeline that performed numerous data cleaning tasks. First, all words were changed to lowercase and the punctuation was removed. Next, all stop words were removed from each essay. Stop words are common words used throughout the essays and carry little information

about the topics being used (Wilbur & Sirotkin, 1992). For example, common stop words seen in essays written by students include *a, are, can, so, the, will,* and *you*. Stop words, if not removed, tend to dominate the estimated topics and reduce the interpretability of the model (Choi et al., 2017). Finally, all words were stemmed and corrected for spelling. Stemming retrieves the base word from the different variations of the same word due to the different tenses or grammatical number. The stemming process increases the clarity of the topics and the interpretability of the model (Schofield et al., 2017). These data cleaning tasks provide clearer results and are essential for fitting topic models.

The sLDA model requires the number of latent topics to be determined a priori. Selecting an appropriate number of topics is nontrivial. Common selection techniques for the number of topics are perplexity or the accuracy of a downstream task (D. M. Blei et al., 2003), the deviance information criterion (DIC; Spiegelhalter et al., 2002), and the Watanabe-Akaike information criterion (WAIC; Watanabe & Opper, 2010). Wheeler, Cohen, et al. (2021) used a simulation study to show the appropriate number of topics with differing amounts of data responses. Given the number of essays and essay lengths of our sample, a three-topic model could be estimated and provide stable measures. Thus, following recommendations from Wheeler, Cohen, et al. (2021), along with comparing DIC values, a three-topic sLDA model was chosen for the analysis in this study.

### 3.4.5 Analysis Plan

The accuracy ratings were used to create a latent continuum with the HCAM for defining rater accuracy and difficulty-to-score of essays. The set of essays were analyzed with a sLDA model to estimate the latent topics and their relationship to the latent continuum defined by the HCAM. In order to better understand the latent topics, a qualitative analysis of the essays was performed by reading individual essays that primarily used one of the three topics. It is important to note that latent topics are similar to latent factors in factor analysis,

and that the term latent topics is not directly related to traditional definitions of topics in the writing assessment literature. Additionally, the source articles – Passage A and Passage B provided by the extended-response item were analyzed with the fitted three-topic sLDA model to assist in defining the latent topics. Finally, the topic proportions for each essay were used in a multivariate regression with the essays' unfolding locations as the dependent variable to determine which semantic features could explain why essays are more or less difficult to score accurately.

## 3.5 Results

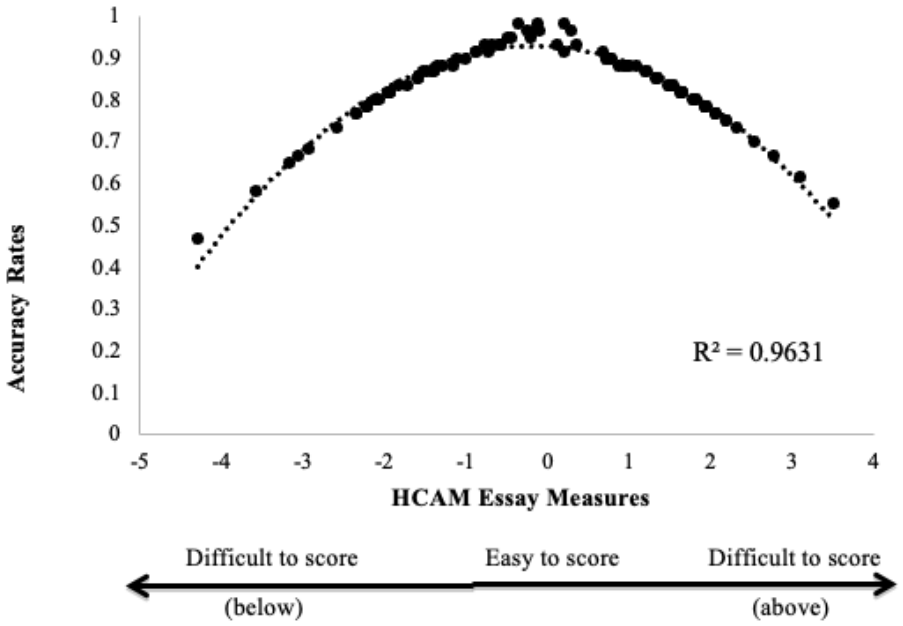
The primary purposes of this study are to identify the latent topics used by students when responding to this particular extended-response item and to identify which of these topics explain the difficult-to-score essays on the unfolding continuum.

### 3.5.1 Unfolding Groups

The 100 essays are centered at zero on the unfolding scale with a standard deviation of 1.72. The essay locations range from $-4.30$ to 3.50. There is a polynomial relationship between accuracy rates and unfolding locations for essays. A higher accuracy rate shows that an essay was easier for raters to score accurately. Therefore, a closer-to-zero unfolding location measure indicates easier-to-score, and a more extreme unfolding location measure implies more difficult-to-score (Figure 3.2).

The unfolding continuum differentiated difficult-to-score essays into two directions, and we named them as difficult-to-score below and difficult-to-score above. We further separated the essays into three groups by their unfolding locations (Figure 3.3). It is somewhat easier to see the effects related to the topics when the essays are categorized with respect to the unfolding continuum. The difficult-to-score below essays ($n = 24$) were defined as having an

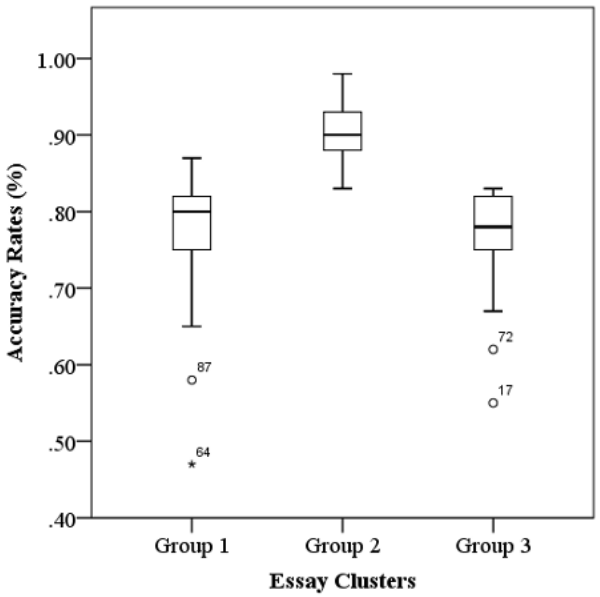
Figure 3.2: Relationship between hyperbolic cosine accuracy model (HCAM) essay measures and observed accuracy rates




*Note.* The accuracy rate is calculated using the sum of accuracy ratings for an essay by all raters divided by the maximum possible points. A second-order polynomial curve explains 96.31% variation in this relationship.

unfolding location less than $-1.5$. Essays in the difficult-to-score below group had accuracy rates between 0.45 to 0.80, indicating that there was discrepancy between the scores given by the operational raters and the scores given by expert raters. The easy-to-score essays ($n = 50$) had an unfolding location between $-1.5$ and 1.5. Essays in the easy-to-score groups had accuracy rates between 0.80 and 0.98, indicating that there was little discrepancy between operational and expert ratings. The difficult-to-score above essays ($n = 26$) were categorized as having an unfolding location greater than 1.5. Similar to the difficult-to-score below group, essays in the difficult-to-score above group had accuracy rates between 0.50 and 0.80, indicating discrepancy between operation and expert ratings.

Figure 3.3: Distribution of accuracy rates within each essay cluster




*Note.* Group 1 - difficult-to-score below, Group 2 - easy-to-score, and Group 3 - difficult-to-score above.

### 3.5.2 Latent Topics and Topic Proportions

The latent topics help define the unfolding continuum, and explore what students are writing about that causes their essays to be difficult to score differently. Particularly, the sLDA models each topic and estimates the probability of each word occurring. By looking at the 25 words with the highest probability for each topic, we get a better sense of the substantive meaning of each topic. Table 3.1 shows the top 25 words for each of the three topics estimated in the sLDA model with estimated probability of that word occurring under the given topic. It is important to note that although the probabilities of the word occurring look small, they are relatively large. For instance, since the total number of unique words across all essays is 421, if all words under a topic were equally likely to occur, the probability would be $\frac{1}{421} = 0.002$. This means a word with a probability of 0.012 is 6 times more likely to occur

than random. Therefore, Table 1 also provides a probability ratio to express how much more likely a word is to occur compared to a random selection from a uniform distribution.

Table 3.1: The top 25 words and their probabilities of occurring for Topic 1, Topic 2, and Topic 3

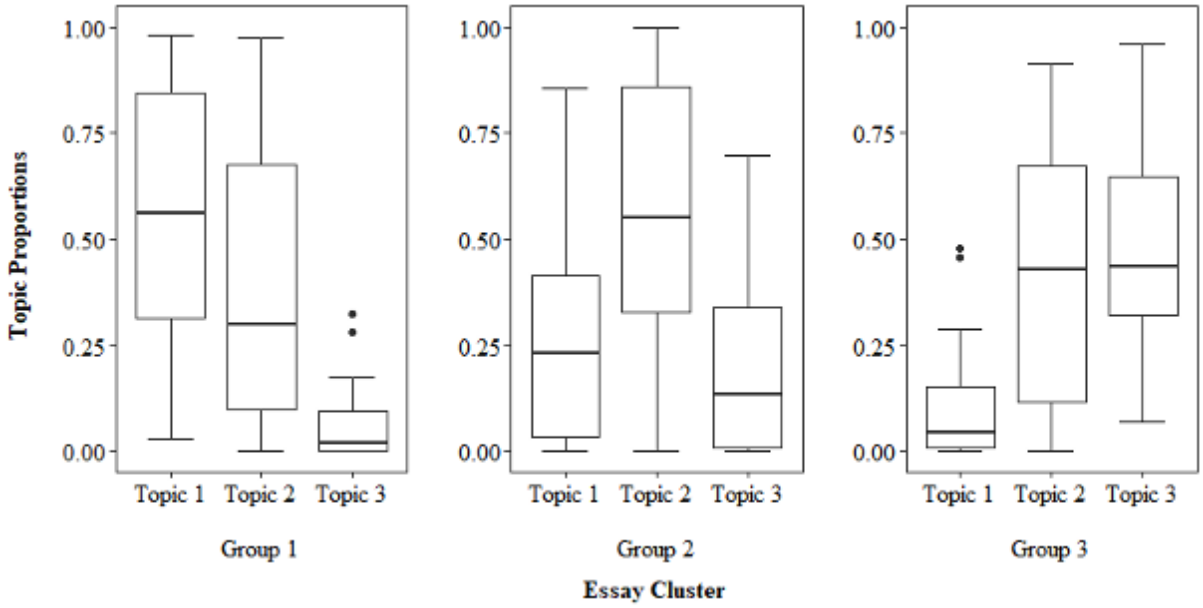
| Topic 1 | | | Topic 2 | | | Topic 3 | | |
|---|---|---|---|---|---|---|---|---|
| Word | Probability | Probability Ratio | Word | Probability | Probability Ratio | Word | Probability | Probability Ratio |
| water | 0.110 | 46.31 | water | 0.097 | 40.84 | water | 0.104 | 43.78 |
| africa | 0.029 | 12.21 | school | 0.028 | 11.79 | oil | 0.030 | 12.63 |
| many | 0.026 | 10.95 | children | 0.025 | 10.53 | africa | 0.020 | 8.42 |
| people | 0.024 | 10.10 | africa | 0.024 | 10.10 | people | 0.019 | 8.00 |
| oil | 0.021 | 8.84 | country | 0.021 | 8.84 | crisis | 0.014 | 5.89 |
| drink | 0.019 | 8.00 | because | 0.021 | 8.84 | supply | 0.014 | 5.89 |
| pollute | 0.015 | 6.32 | ease | 0.019 | 8.00 | contaminate | 0.014 | 5.89 |
| because | 0.014 | 5.89 | time | 0.017 | 7.16 | more | 0.012 | 5.05 |
| crisis | 0.012 | 5.05 | unclean | 0.016 | 6.74 | die | 0.012 | 5.05 |
| waste | 0.012 | 5.05 | crisis | 0.014 | 5.89 | cause | 0.011 | 4.63 |
| children | 0.012 | 5.05 | people | 0.014 | 5.89 | spill | 0.011 | 4.63 |
| contaminate | 0.012 | 5.05 | state | 0.013 | 5.47 | was | 0.010 | 4.21 |
| country | 0.012 | 5.05 | drink | 0.012 | 5.05 | passage | 0.010 | 4.21 |
| these | 0.012 | 5.05 | cause | 0.012 | 5.05 | most | 0.009 | 3.79 |
| make | 0.010 | 4.21 | clean | 0.011 | 4.63 | paragraph | 0.009 | 3.79 |
| animal | 0.009 | 3.79 | consequence | 0.011 | 4.63 | hole | 0.009 | 3.79 |
| need | 0.009 | 3.79 | many | 0.011 | 4.63 | walk | 0.008 | 3.37 |
| sick | 0.009 | 3.79 | her | 0.010 | 4.21 | etana | 0.008 | 3.37 |
| clean | 0.009 | 3.79 | etana | 0.010 | 4.21 | reason | 0.008 | 3.37 |
| find | 0.009 | 3.79 | passage b | 0.010 | 4.21 | pollute | 0.008 | 3.37 |
| consequence | 0.008 | 3.37 | more | 0.009 | 3.79 | passage a | 0.008 | 3.37 |
| mile | 0.008 | 3.37 | passage a | 0.009 | 3.79 | effect | 0.008 | 3.37 |
| day | 0.008 | 3.37 | miss | 0.009 | 3.79 | bug | 0.007 | 2.95 |
| dump | 0.008 | 3.37 | these | 0.009 | 3.79 | day | 0.007 | 2.95 |
| kill | 0.008 | 3.37 | day | 0.009 | 3.79 | many | 0.007 | 2.95 |

*Note.* A uniform distribution over words would carry a probability of 0.002, in which case majority of the top 25 words are more likely to occur by a factor larger than 4. The Probability Ratio column indicates how much more likely each word is to occur compared to a random uniform distribution.

The three topics estimated from the sLDA model have quite a bit of overlap, for example, the word water appears at the top of each topic. This is expected with extended-response items since the answers are constrained by the prompt (Choi et al., 2017). The sLDA model, however, is able to pick up on the subtle nuances of how these words are being used by the student and is able to distinguish between the topics. By doing so, the model provides topic proportions for each essay, which represents the usage of each topic in each essay. The sLDA model estimated that the average topic proportion across all essays for Topic One was 0.289 (0.287), the average topic proportion for Topic Two was 0.474 (0.331), and the average topic proportion for Topic Three was 0.237 (0.242). From this result, we see that overall Topic Two was used the most, followed by Topic One, and then Topic Three.

The analysis of topic proportions can be further broken down into the unfolding groups (Figure 3.4). From this view, we see that the difficult-to-score below group (Group 1) has a high average proportion of Topic One and a low average proportion of Topic Three. This means that the essays from the difficult-to-score below group mainly used words of Topic One. Similarly, the easy-to-score group (Group 2) and the difficult-to-score above group (Group 3) have a high average Topic Two and Topic Three proportion, respectively. This means that the easy-to-score group is mainly using words of Topic Two while the essays of the difficult-to-score above group is mainly using words in Topic Three.

Figure 3.4: Box plots for the usage of each topic with respect to the unfolding groups




*Note.* The sum of the topic proportions for each essay equals 1. Group 1 - difficult-to-score below, Group 2 - easy-to-score, and Group 3 - difficult-to-score above.

### 3.5.3 sLDA Regression Results

Along with the estimated topics and topic proportions, the sLDA model captured the relationship between the unfolding locations and the topics through a regression analysis. Each essay was considered as an observation; thus, our regression analysis had 100 observations

($n = 100$). The unfolding location for each essay was the dependent variable and ranged between $-4.3$ and 3.5. The topic proportions for the three topics for each essay were our independent variables and ranged between 0 and 1. Table 3.2 provides the sLDA regression results. The regression model showed that the topic proportions could moderately explain the variation in the unfolding location ($r^2 = 0.561$) and found that Topics One and Three had a significant linear relationship with unfolding location measures ($p < 0.001$) while Topic Two did not have a significant linear effect ($p = 0.515$). Furthermore, the regression coefficient for Topic One was negative ($\beta_{\text{Topic 1}} = -2.712$), indicating that the difficult-to-score below group have higher proportions on Topic One. The regression coefficient for Topic Three was positive ($\beta_{\text{Topic 3}} = 3.593$), indicating that the difficult-to-score above group have higher proportions on Topic Three. Although Topic Two did not show a significant linear effect on the unfolding measures, it may have other types of relationship (e.g., polynomial) with the unfolding continuum.

Table 3.2: sLDA Regression Analysis: Topic Proportion Effects on Unfolding Location

| Effect | Estimate | SE | 95% Confidence Interval | | p-value |
|---|---|---|---|---|---|
| | | | Lower Limit | Upper Limit | |
| Topic One | -2.712 | 0.313 | -3.339 | -2.087 | <.001 |
| Topic Two | -0.143 | 0.218 | -0.579 | 0.293 | 0.515 |
| Topic Three | 3.593 | 0.385 | 2.823 | 4.363 | <.001 |

*Note.* Independent variables were the topic proportions of all three topics for each essay. Dependent variable was the unfolding location measures. Topic proportion values ranged from 0 to 1. Unfolding location values ranged from -4 to 4.

The relationship between the topics and the unfolding groups is easier to see through box plots shown in Figure 4. That is, Topic One was mainly used by the difficult-to-score below group, Topic Two was mainly used by the easy-to-score group, and Topic Three was mainly used by the difficult-to-score above group.

### 3.5.4 Defining Latent Topics

The sLDA model was able to distinguish between the topics in each response. In order to further define the topics, a qualitative analysis of each essay was performed. Specifically, we read essays from the three unfolding groups and analyzed the unfolding location, topic proportions, and raw text. Table 3.3 provides three responses from the 100 essays in this study. Student A's essay belongs to the difficult-to-score below group and primarily used Topic One in their response. This essay response did use the evidence from the passages provided, however, it failed to provide specific citations to the corresponding passages. Student B's essay belongs to the easy-to-score group and primarily used Topic Two in the response. This essay response followed the instruction by using and citing the evidence from both source articles to support its own argument. Student C's essay belongs to the difficult-to-score above group and primarily used Topic Three in their response. This essay response cited the examples from one of the passages provided, namely, Passage A.

To further demonstrate the connection between the textual borrowing feature of an essay and how difficult it is for raters to score accurately, we fit the sLDA model to the source passages provided in the extended-response item. The appendix provides both passages and their estimated topic proportions. The results show that Passage A mainly uses Topic Two ($\tau_{\text{Passage A,Topic 2}} = 0.536$) and Topic Three ($\tau_{\text{Passage A,Topic 3}} = 0.389$) while Passage B mainly uses Topic Two ($\tau_{\text{Passage B,Topic 2}} = 0.874$). In other words, Topic Two is related to both of the passages, Topic Three is related to Passage A only, and Topic One is not necessarily related to either of the passages.

The unfolding continuum differentiated the essays into two directions that are difficult-to-score below and difficult-to-score above. The essays with different unfolding locations used different latent topics in the responses. Table 3.4 shows a meaningful name and description to define each of the three topics. The extended-response item asks the students to use both the passages provided and use two pieces of evidence from each passage. Our results suggest that

Table 3.3: Sample responses with unfolding location and topic proportions

| **Student A's Essay – Difficult-to-Score Below Group (Unfolding Location = −1.969) Topic 1 Proportion = 0.842, Topic 2 Proportion = 0.140, Topic 3 Proportion = 0.018** |
|---|
| In Africa the water is very polluted and dangaroose. Many Africans die every day from illnesses and mostly dehydration. Since all of the fresh water of polluted relly nothing is safe to drink in africa. So any water they do drink its unsafe. When they get sick they have to miss school for a certain amount of time. They stay out until there healthy again. Wich could take a long time. The illneses are very deadly if not if not treated right. Their working on ways to convert a solid into a liqid. A human being can only last 3-4 days without fresh, clean water. Not only that drought occures very often in africa. Sometimes all you can find is a mud hole in the bottom of a dried up water bed. That all the info I could find. |
| **Student B's Essay – Easy-to-Score Group (Unfolding Location = 0.966) Topic 1 Proportion = 0.067, Topic 2 Proportion = 0.933, Topic 3 Proportion = 0** |
| **...** The first conquense I will be talking about is children not being able to go to school. "Because she has to make this trip two more times today, there is no time to go to school." This quote from **passage B** is from when Etona is feticing water from a water hole. It begins to say she has no time to go to school, which means she won't get into college; that means she will not get a job (or at least one that pays much); that will hurt the economy. A statement from **passage A** that further supports that which I am saying is "children in some African countries often miss school because of illnesses resulting from unclean water and poor sanitation practices." This further support my statment saying that the water that taking long periods of time to fetch can make you sick causing you to miss more days of school. This brings me into my second conquence.<br>Secondly, people become sick from unclean water. My reasoning for this is this statement from **passage A**. "More than 85% of all dieseses in African children are caused by unclean water." This states that nearly all of dieseses children can get are result of unclean water. This will lead to a much higher mortality rate as well. I can support this with these statements from **passage B**. "She has known many who have become sick from drinking polluted water. Some of her friends have died from dieseses they contracted from the polluted water." This statement (as prevrously said) supports my idea at large amonts of sicknesses and deaths from those sicknesses in Africa... |
| **Student C's Essay – Difficult-to-Score Above Group (Unfolding Location = 2.314) Topic 1 Proportion = 0.153, Topic 2 Proportion = 0.106, Topic 3 Proportion = 0.741** |
| ...In some parts of Africa, the urbanization and oil spills are some factors that happen to contribute to the water crisis. "As more and more people move closer to populated cities, the food supply demand grows. Farming and agriculture have increased, creating the need for more fertilizers that are damaging the water supply." This piece of textual evidence from **Passage A** is one cause for the countries in Africa to have low water supply. Sadly, that is not the only cause. "Nigeria is Africa's largest oil producer. Last year, 6,000 tons of oil were dumped into the Niger Delta waterway." This textual evidence from **Passage A** is another cause to add to the list. Since the oil just sits on top of the water, it kills and contaminates alot of the plants and animals. The animals don't really have a choice to drinking it or not because there is not much water source around them, so they have to drink what they have. One more consequence of the water crisis is that a person can only live for so long without water. The estimated amount of days are 3-4. Another consequence is that alot of Africans have to walk several miles each day in order to get the water to support their family. "Her village has no running water, so Etana and the other women and children must make a daily trek to get water." This shows that many people have to walk to get water. "She is on her way to the only water source she knows, which is located several miles away." This shows that they had to walk several miles... |

*Note.* One sample essay from each of the unfolding location groups. All passage references are bolded.

Topic One is for students who used evidence in the passages but failed to cite their sources. Failing to cite the sources caused the essays to be difficult-to-score accurately because the

Table 3.4: Topic names and descriptions

| Topic | Name | Description |
|---|---|---|
| Topic One | Provided Evidence without Citing Source | Students who used Topic One provided evidence but did not follow the prompt by not citing either of the passages provided |
| Topic Two | Provided Evidence with Citing Source (Passage A and Passage B) | Students who used Topic Two provided sufficient evidence and followed the prompt by citing both passages provided |
| Topic Three | Provided Evidence with Citing Source (Passage A) | Students who used Topic Three provided evidence but did not fully follow the prompt by citing only one of the passages provided |

student answered the prompt but did not completely follow instructions on composition. Topic Two is for students who used evidence in both passages and cited their sources. Since the students answered the prompt and followed the instructions, their essays were easy-to-score accurately. Topic Three is for students who used evidence from Passage A and cited their source, but failed to use both passages. Failing to use both passages caused essays to be difficult-to-score accurately because the student answered the prompt and provided evidence but did not completely follow instructions.

## 3.6 Discussion

This study proposed the use of a sLDA model to empirically define the substantive meaning of an accuracy unfolding continuum. Student essays and raters were placed onto the common unfolding scale. The location measures were obtained based on the hyperbolic cosine accuracy model with the use of accuracy ratings. Results based on the sLDA regression analysis and

box plots showed that all three of the topics have a moderate influence on the unfolding location measures. Specifically, Topic One is shown to be associated with essays that are difficult-to-score below, that is, student essays that mainly consisted of words in Topic One were difficult to score accurately. Topic Two is shown to be associated with essays that are easy-to-score, that is, the essays that used more of Topic Two were easy to score accurately. Topic Three is shown to be associated with essays that are difficult-to-score above, that is, these essays were also difficult to score accurately.

A qualitative analysis of the essays and the passages provided from the extended-response item further defined the latent topics. The easy-to-score essays generally had highest topic proportion on Topic Two, and the responses strictly followed the instruction that (a) cite at least two examples from Passage A, (b) cite at least two examples from Passage B, and (c) explain how these examples illustrate the discussed topic (i.e., consequences of water crisis). The essays of the difficult-to-score below group typically had highest topic proportion on Topic One. These essay responses used the examples but failed to cite the passages provided. On the other hand, the essays of the difficult-to-score above group had highest topic proportion on Topic Three, and the responses cited examples from mainly Passage A.

It is worth noting that easy-to-score essays are those essays that raters possess more consistent and accurate judgments about the reflected level of writing proficiency. On the contrary, when raters possess more inconsistent and inaccurate judgments toward an essay, the essay becomes more difficult-to-score. Based on the accuracy ratings, the unfolding accuracy continuum shows how difficult for raters to score an essay accurately, rather than the level of writing proficiency of each student.

The results from the topic models indicate interpretation for this particular writing assessment. That is, the estimated sLDA model is dependent on the prompt. This is important to note since it limits the generalizability of the results. The analysis, however, suggested one potential, generalizable reason why an essay is either difficult-to-score accurately or easy-to-

score accurately based on the content topics. J. Wang et al. (2017) explored rater perceptions during the scoring of an integrated writing assessment and found that raters did not have the same understanding of the appropriate amount of textual borrowing and ways of citing the evidence from source articles. This supports our findings of the relationship between latent topics and the unfolding accuracy continuum. The latent topics reflect different amount of textual borrowing as well as the citing sources. Inconsistent perception among raters toward this essay feature is identified as a source of inaccuracy.

This study provides an approach based on unfolding models and sLDA models for exploring different reasons that some essays are more difficult to score accurately than others. Our findings can help researchers, educators, and classroom teachers in writing assessments better understand the relationship between essay characteristics and rater scoring accuracy, which will further identify sources of inaccuracy and improve the rating quality.

# Chapter 4

# Textual Data as Process Data: A New Scoring Procedure to Improve Ability Estimation for Mixed-Format Assessments[3]

[3] Wheeler, J. M., Wang, S., Tan, Y., & Cohen, A. S. Submitted to *Psychometrika*, 7/27/2022

## 4.1 Preface

The study presented in this chapter is currently under review in Wheeler, Wang, Tan, and Cohen (under revision). The study was directed under the supervision of Drs. Shiyu Wang and Allan S. Cohen. My role in this study was to help develop the adaptation of a process-data-based scoring procedure to mixed-format assessments, develop the bootstrapping script that implemented the proposed scoring procedure, and draft the initial manuscript. We plan to conduct a follow-up simulation study to generalize the results. This chapter demonstrates how features extracted from topic models can be incorporated into IRT scoring for mixed-format assessments.

## 4.2 Abstract

Process data in educational measurements refer to the intermediate data that capture the interaction between an examinee and an item on an assessment, which describe the answering behavior of the examinee. With the expansion of technology in educational measurement, the collection of process data has grown substantially. More recently, studies have been conducted on how to incorporate these process data into the ability estimation of examinees. This study expands on previous studies by extracting process features from textual responses to constructed-response items using latent semantic analysis (LSA) and latent Dirichlet allocation (LDA). Furthermore, the information extracted using LSA and LDA is then used as process data to improve the ability estimates of examinees. The results of this study suggest that the use of the proposed scoring procedure for mixed-format assessments provides more accurate ability estimates than traditional item response theory scoring.

## 4.3 Introduction

Mixed-format assessments, which contain both multiple-choice (MC) and constructed-response (CR) items, are commonly used for achievement testing. MC items are desirable for evaluations due to their efficiency in scoring and their ability to provide more objective and reliable measures than CR items (S. Kim et al., 2010). On the other hand, CR items often reflect real-world tasks and, therefore, are thought to assess high-order ability and lead to minimal construct underrepresentation (Messick, 1994; Weigle et al., 2013). Mixed-format assessments have recently gained popularity due to the ability to score CR items using machine learning algorithms, thus eliminating the major drawback of the cost and time to score CR items (Shermis, 2014). Due to the benefits of MC and CR items, along with the ability to machine score CR items, mixed-format assessments are widely used within large-scale assessment systems that utilize computer-based assessment platforms. The final score for each ability is typically obtained via a two-step procedure. First, a categorical score is generated using either a machine score approach or based on human rating following certain rubrics. Second, an item response theory (IRT) model is selected to fit the categorical response and provide an estimate for each examinee and thus establish a score scale based on it.

This two-step scoring procedure within an IRT framework has some limitations. A major concern in this regard is that the final ability score is obtained based only on the categorical score. That is, once the scores are assigned, the actual text is no longer considered for ability estimation. A conjecture in this study is that we may lose some useful information from the original textual response to infer the latent abilities of an examinee. In fact, textual responses to CR items can also be considered as a type of process data because they directly reflect an examinee's process of responding to CR items. Process data, in general, refer to various sorts of data collected during examinees' process of responding, including things such as reaction time, eye-movement, and sequence of actions performed. Recent research

studies in educational measurement have revealed that process data can contain additional useful information about examinees that are not found in response patterns alone (e.g., Cho et al., 2020; Molenaar et al., 2016). Moreover, recent studies have shown that additional information about examinees can be extracted from CR items through topic models (He et al., 2019; S. Kim et al., 2017).

Zhang et al. (2021), in a recent study, proposed a new item response theory scoring procedure that incorporates process features extracted from process data. The study by Zhang et al (2021) demonstrates how process features can be used to improve the accuracy and reliability of ability estimates compared to the traditional IRT scoring procedure. The items in that study, however, did not include mixed-formats and specifically looked at process features extracted from log-file data. There has not yet been a study reported that proposes a scoring procedure framework for mixed-format assessments that utilizes features extracted from CR items. This is unfortunate, since, as is suggested above, there is evidence that information in the text of examinees' answers does appear to be related to but not included in the standard IRT estimation of ability. Therefore, this study adapts the procedure proposed by Zhang et al. (2021) and introduces a new scoring procedure framework specifically for mixed-format assessments.

The remainder of this paper is structured as follows: Section 2 describes the process-based scoring procedure proposed by Zhang et al. (2021). In Section 3, the new scoring procedure for mixed-format assessments is introduced by adapting the process-based scoring procedure. Section 4 illustrates the new scoring procedure through an empirical case study and investigates our research questions. Section 5 concludes the paper with a discussion about practical considerations of the proposed scoring procedure and directions for future research on scoring mixed-format assessments.

## 4.4 Review and Background

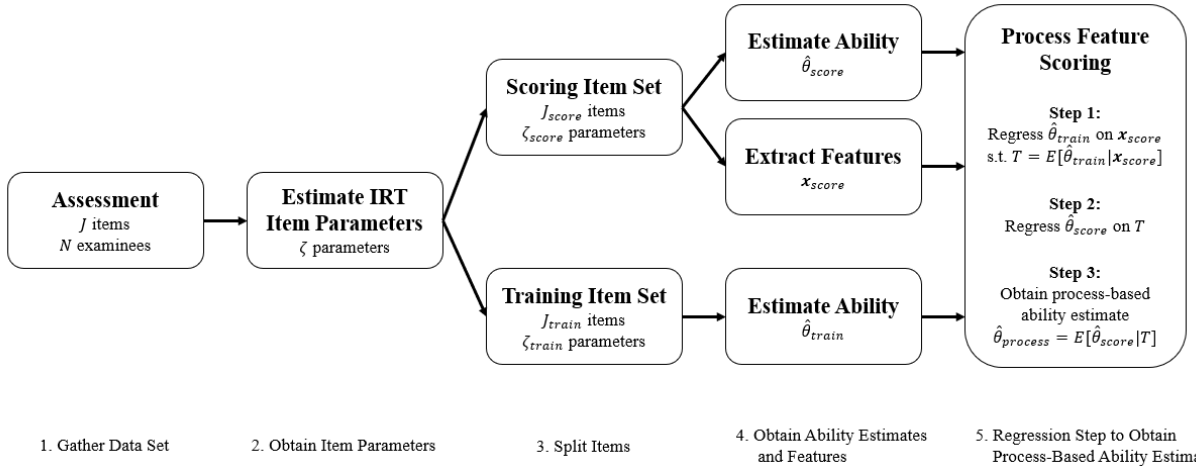
In recent years, the field of measurement has seen a rise in the development of scoring procedures that incorporate process features (e.g., Ercikan & Pellegrino, 2017; Gobert et al., 2013). Additionally, with computerized assessments being more abundant, the use of mathematical models to analyze textual data has become increasingly popular. This section reviews a process data scoring procedure and two text analysis models that have been previously used in educational measurement. Later in the present study, these methods are augmented to create a new scoring procedure for mixed-format assessments.

### 4.4.1 A IRT Scoring Procedure using Process Data

Zhang et al. (2021) proposed a multistage approach to incorporate process data in the IRT-based scoring algorithm. This scoring procedure assumes that an assessment designed to measure a single latent trait, $\theta$, contains $J$ items. Each item has a set of item parameters $\boldsymbol{\zeta}$ and is given to $N$ examinees. The goal of the multistage approach is to provide a single $\theta$ estimation for each examinee using both response data and process data from computer log files to the $J$ items.

Figure 4.1 documents the key steps of this procedure. In the first stage, response data are collected from the given assessment. The second stage estimates the item parameters using an IRT model along with the response data that were collected. In the third stage, the items are split into two distinct sets, namely, the training set and the scoring set. The fourth stage estimates two ability parameters for each examinee using the training and scoring item sets. Specifically, each examinee is given two separate ability estimates based on their responses to the items in the training and scoring sets, namely, $\hat{\theta}_{train}$ and $\hat{\theta}_{score}$), respectively. Additionally, process features are extracted from the log file data to the items in the scoring set for each examinee, $\mathbf{x}_{score}$. The final stage uses the estimated abilities and the process

Figure 4.1: Scoring procedure that incorporates the latent features from process data into IRT ability estimation




features to calculate a final ability estimate, $\hat{\theta}_{process}$, for each examinee using a regression approach. The final ability estimate is based on the process data and is referred to as the process-based ability estimate for an examinee.

Zhang et al. (2021) established the theoretical performance gain of this IRT-based process data scoring procedure over traditional IRT scoring procedures, given a series of assumptions and conditions. The performance gain of the scoring procedure was demonstrated using an empirical example from an assessment consisting of dichotomous MC responses and log-file process data. Process features were extracted from the log file data using multidimensional scaling (MDS; Tang et al., 2020) and the scoring procedure was applied to the empirical data. The results showed that the process data scoring procedure provided more accurate ability estimates than traditional IRT scoring procedures.

The scoring procedure proposed by Zhang et al. (2021) provides a general framework that suggests the potential to adapt the method to scoring mixed-format assessments. First, as illustrated in the Introduction section, each CR item is associated with two types of

data: the original textual response data, and the categorical score generated based on the textual response using a machine score or rubric score mechanism. This implies that the textual response data contain at least as much information as the final outcomes, which is an assumption imposed by Zhang et al. (2021). Second, the mixed-format assessment for this study satisfies the A1 (monotonicity) and A2 (local independence) assumptions, as described in Zhang et al. (2021), because the scoring of mixed-fromat assessments uses an IRT framework, which automatically guarantees the satisfaction of these two assumptions.

The unique structure of mixed-format assessments, however, impose several challenges for applying their method directly. First, in the Zhang et al. (2021) study, the response data consisted of only dichotomous responses. Mixed-format assessments often contain items that are scored both dichotomously and polytomously; therefore, implementing these scoring procedures for mixed-item response types still needs to be properly investigated. Secondly, the process features used in their study were extracted from each item. That is, each item had its own set of process features based on the sequence of actions an examinee made while responding. For mixed-format assessments, however, a type of process data is from the textual responses to CR items. This means that only a subset of items contain process data rather than all items. Lastly, the splitting of the items during their study was entirely random. This means that there are no special considerations about which items should belong to the training and scoring sets. For mixed-format assessments, since only some of the items have process features (i.e., the CR items), splitting the items into training and scoring sets is somewhat limited. Furthermore, the statistical item information of items may differ drastically due to the different response types in a mixed-format assessment; therefore, implementing additional constraints on the training and scoring sets may improve the overall performance of the scoring procedure, which was not investigated.

### 4.4.2 Feature Extraction Methods for Textual Data

We suggest that these textual responses to CR items can be seen as a type of process data since they provide information about examinees' problem-solving process (He & Davier, 2015). Similar to other response process data, such as log-file data, the textual responses to CR items are complex in that each response is unique and the length varies across examinees, meaning a feature extraction method must be employed to obtain useful information from the textual responses. Features are commonly extracted from CR items using various topic models.

Topic models are a family of mathematical and statistical techniques used to analyze textual data. In educational measurement, topic models are often used to extract latent features, referred to as topics, from CR items. Two topic models that are commonly used in measurement are latent semantic analysis (LSA; Deerwester et al., 1990) and latent Dirichlet allocation (LDA; D. M. Blei et al., 2003). The LSA model is often used as the basis of automated scoring engines for CR items (e.g., Foltz et al., 1999; Graesser et al., 2004). These scoring algorithms focus on assigning categorical ratings to each CR item, rather than estimating an ability for the entire assessment. LSA is a popular and well studied method for text analysis as it often provides scores to essays in high agreement with human-provided scores (Landauer et al., 1998).

The LDA model has relatively new applications within measurement. LDA is often used as a post hoc analysis to provide additional information about the written responses to CR items. The results from the LDA model are clusters of words that provide information about common topics that examinees use when responding to CR items. Unlike LSA, LDA is typically not used as the basis for automated scoring engines because the topics are not necessarily related to the scores assigned and the results are based on exploratory approaches (S. Kim et al., 2017; Wheeler, Engelhard, et al., 2022a). A benefit of LDA, however, is that it

can identify transitions in writing after an instructional intervention that are not necessarily reflected in the overall scores (Duong et al., 2019; S. Kim et al., 2017).

**Latent Semantic Analysis (LSA).**

LSA uses a matrix decomposition technique known as singular value decomposition (SVD) that extracts features from textual data. LSA is related to principal component analysis (PCA) and extracts latent components found in a set of answers to CR items. Similar to component scores, LSA also extracts features for each examinee that express the usage of the latent components. Suppose that a set of responses to a CR item has $D$ documents where each document, $d$, has $N^{(d)}$ words, and suppose that there are $W$ unique words across all documents. LSA first constructs a document-by-word matrix $\boldsymbol{Y}_{(D,W)}$ where the rows are associated with the documents and the columns are associated with the unique words. Each element in the document-by-word matrix, $y_{d,w}$, contains the number of times the $w^{th}$ word appears in the $d^{th}$ document. LSA then applies the SVD technique to the document-by-word matrix to obtain three orthonormal matrices: a document matrix $\boldsymbol{D}_{(D,K)}$, a word matrix $\boldsymbol{W}_{(W,K)}$, and a diagonal topic matrix $\boldsymbol{\Sigma}_{(K,K)}$. Multiplying the three orthonormal matrices together results in the original document-by-words as shown in

$$\boldsymbol{Y}_{(D,W)} = \boldsymbol{D}_{(D,K)}\boldsymbol{\Sigma}_{(K,K)}(\boldsymbol{W}^T)_{(K,W)}, \tag{4.1}$$

where $K$ is the full rank of the document-by-word matrix such that $K = min(D, W)$ and $(\boldsymbol{W}^T)_{K,W}$ is the transpose of the word matrix $\boldsymbol{W}_{W,K}$. The word matrix contains the $K$ latent components found in the set of textual responses to the CR item, the document matrix contains the component scores of each response for the $K$ components, and the topic matrix contains the component weights, known as singular values (similar to eigenvalues), for each of the $K$ components.

A subset of $M$ components is selected based on the singular values such that $M \leq K$. There are several methods for selecting an appropriate number of components, including fit indices, variance-explained metrics, cross-validation methods, and performance-based methods of downstream tasks (Wild, 2016). Once the appropriate number of components is selected, the document, word, and topic matrices are reduced by removing the components not selected. The resulting matrices are referred to as the reduced document matrix $\boldsymbol{D_r}_{(D,M)}$, reduced word matrix $\boldsymbol{W_r}_{(W,M)}$, and reduced topic matrix $\boldsymbol{\Sigma_r}_{(M,M)}$. The process features are calculated by multiplying the reduced document matrix by the reduced topic matrix,

$$\boldsymbol{X}_{(D,M)} = \boldsymbol{D_r}_{(D,M)} \boldsymbol{\Sigma_r}_{(M,M)}, \tag{4.2}$$

where $\boldsymbol{X}_{D,M}$ is the process feature matrix, which contains the weighted component scores for each textual response. Each row of the process feature matrix contains the set of features for the textual response associated with that row. Therefore, a set of process features for document $d$ is defined as

$$\vec{x}_d = \boldsymbol{X}_{d,*} = [x_{d,1}, x_{d,2}, ..., x_{d,M}] \tag{4.3}$$

**Latent Dirichlet Allocation (LDA).**

The LDA model estimates three main parameters from a corpus of documents: a set of word probabilities (i.e., topics) $\boldsymbol{\beta}$, a set of topic proportions for individual documents $\boldsymbol{\theta}$, and a set of topic assignments $\boldsymbol{z}$. Each topic is a set of probabilities over the vocabulary $\boldsymbol{V}$, which is the set of unique words found across all documents. The set of probabilities that constitute a topic expresses the probabilities of each word from the vocabulary appearing under the given topic. Each document is given a set of topic proportions, which is a vector of proportions over the topics. The topic proportions express the proportion of each topic used in the given

document. Additionally, each document is given a set of topic assignments that represent the topic membership of each word that appears in the document.

The LDA model is a hierarchical mixture model where topics are considered corpus-wide parameters, topic proportions are considered document-wide parameters, and topic assignments are considered word-wide parameters. The mixture component of LDA is that topics are a mixture of words from the vocabulary, and topic proportions are a mixture of topics. Additionally, LDA assumes the following generative process: (1) assume that there are $K$ topics, (2) assume each document is generated by first generating its topic proportion, (3) assume that a topic is assigned to each word in the document which is determined by its topic proportion, and (4) assume each word in the document is generated given the topic assignment and topic distribution.

Suppose that there is a corpus that contains $d = 1, 2, \ldots, D$ documents, each document contains $N_d$ words, and there are $V$ unique words across all documents. The LDA model assumes a priori that there are $K$ corpus-wide topics. The joint distribution for the observed word variables ($\boldsymbol{w}_{1:D}$), the latent topic assignment variables ($\boldsymbol{z}_{1:D}$), the topics ($\boldsymbol{\beta}_{1:K}$), and the topic proportions ($\theta_{1:D}$) is given by

$$p(\boldsymbol{w}_{1:D}, \boldsymbol{z}_{1:D}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\theta}_{1:D}|\boldsymbol{\eta}, \boldsymbol{\nu}), \tag{4.4}$$

where $\boldsymbol{\eta}$ and $\boldsymbol{\nu}$ are the prior hyperparameters. The joint distribution for LDA can subsequently be factorized into the conditional distributions (i.e., likelihood of the data) and priors, which is shown in

$$\begin{aligned} &p(\boldsymbol{w}_{1:D}, \boldsymbol{z}_{1:D}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\theta}_{1:D}|\boldsymbol{\eta}, \boldsymbol{\nu}) = \\ &p(\boldsymbol{w}_{1:D}|\boldsymbol{z}_{1:D}, \boldsymbol{\beta}_{1:K})p(\boldsymbol{z}_{1:D}|\boldsymbol{\theta}_{1:D})p(\boldsymbol{\beta}_{1:K}|\boldsymbol{\nu})p(\boldsymbol{\theta}_{1:D}|\boldsymbol{\eta}), \end{aligned} \tag{4.5}$$

where $p(\boldsymbol{w}_{1:D}|\boldsymbol{z}_{1:D}, \boldsymbol{\beta}_{1:K})$ is the conditional distribution of the observed words given the topic assignments and the topics and is assumed to follow a multinomial distribution; $p(\boldsymbol{z}_{1:D}|\boldsymbol{\theta}_{1:D})$ is the conditional distribution of topic assignments given the topic proportions and is assumed

to follow a multinomial distribution; $p(\boldsymbol{\beta}_{1:K}|\boldsymbol{\nu})$ is the prior distribution for the topics where $\boldsymbol{\nu}$ is the prior hyperparameter which controls the density of the word probabilities and is assumed to follow a Dirichlet distribution; and $p(\boldsymbol{\theta}_{1:D}|\boldsymbol{\eta})$ is the prior distribution for the topic proportions where $\boldsymbol{\eta}$ is the prior hyperparameter which controls the density of topic proportions and is assumed to follow a Dirichlet distribution (D. M. Blei et al., 2003; Ponweiser, 2012).

Since the joint distribution is intractable due to its dimensionality, LDA infers the latent parameters $(\boldsymbol{z}_{1:D}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\theta}_{1:D})$ through the conditional distribution of these parameters given the observed words: $p(\boldsymbol{z}_{1:D}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\theta}_{1:D}|\boldsymbol{w}_{1:D})$. This distribution is referred to as the posterior distribution, which is proportional to the joint distribution, given by

$$p(\boldsymbol{z}_{1:D}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\theta}_{1:D}|\boldsymbol{w}_{1:D}) \propto p(\boldsymbol{w}_{1:D}, \boldsymbol{z}_{1:D}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\theta}_{1:D}), \tag{4.6}$$

where $p(\boldsymbol{w}_{1:D}, \boldsymbol{z}_{1:D}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\theta}_{1:D})$ is the joint distribution shown in Equation (2). Given the dimensionality of the corpus, the posterior distribution can be factorized as

$$\begin{aligned} &p(\boldsymbol{z}_{1:D}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\theta}_{1:D}|\boldsymbol{w}_{1:D}) \propto \\ &\prod_{k=1}^{K} P(\boldsymbol{\beta}_k|\boldsymbol{\nu}) \prod_{d=1}^{D} \Bigg( P(\boldsymbol{\theta}_d|\boldsymbol{\eta}) \prod_{n=1}^{N_d} P(w_{d,n}, z_{d,n}|\boldsymbol{\theta}_d, \boldsymbol{\beta}) \Bigg). \end{aligned} \tag{4.7}$$

The posterior distribution of LDA can be estimated through various techniques. The two primary methods for estimating the posterior are by Gibbs sampling and variational inference D. M. Blei et al. (2003). The process features for each document are determined by the estimated topic proportions $\boldsymbol{\theta}_{1:D}$. Therefore, a set of process features for document $d$ is defined as

$$\vec{x}_d = \boldsymbol{\theta}_d = [\theta_{d,1}, \theta_{d,2}, ..., \theta_{d,K}] \tag{4.8}$$

## 4.5 Mixed-Format Assessment Scoring Procedure

In this section, we describe the proposed scoring procedure for mixed-format assessments. The general framework of the proposed scoring procedure for mixed-format assessments follows the scoring procedure introduced by (Zhang et al., 2021). The scoring procedure uses a three-step regression approach to incorporate process-data features into the ability estimates. In this section, the scoring procedure is introduced and various considerations are described for applying process-based scoring to mixed-format assessments.

### 4.5.1 Procedure Overview and Mathematical Formulation

Suppose that an assessment contains $J$ items and is given to $N$ examinees and is designed to measure a single latent trait, $\theta$. The scoring procedure assumes an IRT measurement model such that the probability of a score $k$ on item $j$ is a function of the examinee's latent trait and item parameters, as shown in Equation 4.9,

$$P(Y_j = k|\theta, \boldsymbol{\zeta}_j), \tag{4.9}$$

where $Y_j$ is the observed response for item $j$, $k$ is the assigned score, $\theta$ is the latent trait being measured, and $\boldsymbol{\zeta}_j$ are the item parameters for item $j$. The MC and CR items may have different measurement models (e.g., possibly due to the number of rating categories for the CR items). That is, MC items are often scored dichotomously, correct ($k = 1$) or incorrect ($k = 0$), and CR items are sometimes scored polytomously with multiple rating categories where $k = 0$ is the lowest attainable score and $k = K$ is the highest attainable score. Regardless of the measurement models used for each item, the underlying latent trait measured by the item must be the same for this scoring procedure.

For this scoring procedure, the item parameters are assumed to be known and calibrated before estimating $\theta$. Item parameters may be calibrated using various estimation methods,

including marginal maximum likelihood estimation (MMLE) and Bayesian expected a posteriori (EAP) methods (Baker & Kim, 2004). Once the item parameters are estimated, the $J$ items are split into two distinct subsets of items, namely the training set $S_{train}$ and the scoring set, $S_{score}$. The item parameters for the items in the training and scoring sets are denoted as $\boldsymbol{\zeta}_{train}$ and $\boldsymbol{\zeta}_{score}$, respectively. These sets of item parameters, along with the response pattern of the items in each set, are used to calculate ability estimates for each examinee. That is, $\boldsymbol{\zeta}_{train}$ and the response pattern for the training set, $\boldsymbol{Y}_{train}$, are used to calculate an ability estimate based on the training set items, $\hat{\theta}_{train}$. Similarly, $\boldsymbol{\zeta}_{score}$ and the response pattern for the scoring set, $\boldsymbol{Y}_{score}$, are used to calculate an ability estimate based on the scoring set items, $\hat{\theta}_{score}$.

Extracting process features from the CR items in the $S_{score}$ set is the final step before calculating the process-based ability estimates. The textual responses to the CR items are used to estimate the topic models, and the features from the textual responses are used as the process features. For the proposed scoring procedure, we recommend using the LSA and LDA topic models, as they have been shown to be useful models in educational measurement (Attali & Burstein, 2006; Uto, 2019; Xiong et al., 2021). Once the process features are extracted from the CR items in the $S_{score}$ set, let $\boldsymbol{x_{score}}$ denote the process features for an examinee.

The proposed mixed-format scoring procedure then follows the regression procedure proposed by Zhang et al. (2021) to obtain a $\theta$ estimate that contains information gained from the process features extracted from the text. That is, first regress $\hat{\theta}_{train}$ on the text features $\boldsymbol{X}$ and obtain the expected value for each examinee using the regression model denoted as $T_{process}$. Then, regress $\hat{\theta}_{score}$ on $T_{process}$ using another regression step. Lastly, obtain the expected value from the OLS regression model from the previous step denoted as $\hat{\theta}_{process}$. This estimated $\hat{\theta}_{process}$ is the new estimate of latent ability of an examinee that

incorporates information obtained from the textual responses to the CR items. The proposed scoring procedure can be summarized by the following three steps:

Step 1: Regress $\hat{\theta}_{train}$ on $\boldsymbol{x_{score}}$ such that $T_{process} = E[\hat{\theta}_{train}|\boldsymbol{x_{score}}]$

Step 2: Regress $\hat{\theta}_{score}$ on $T_{process}$

Step 3: Obtain the process-based ability estimate $\hat{\theta}_{process} = E[\hat{\theta}_{score}|T_{process}]$

### 4.5.2 Issues Related to Mixed-Format Assessments

As mentioned in Section 2.1, applying the proposed mixed-format assessment scoring procedure described above poses several challenges due to the mixed-item types and the restriction of process features being extracted from only the CR items.

**Issues related to extracting features using topic models.**

There are several considerations when extracting features from CR items using LSA and LDA. The features extracted from LSA are based upon the deconstruction of the document-by-word matrix; therefore, there is no assumed underlying statistical model. This means that LSA can handle various data sizes and still extract meaningful features (CITE SOURCE). This also implies that we can expect that the features produced from LSA preserve the full information on the final response, which is an important assumption imposed by Zhang et al. (2021). The LDA model, however, has an assumed statistical model, which means the reliability of the topic proportion estimates depends on the amount of data used in the estimation algorithm. That is, the process features extracted using LDA are estimates and contain measurement error, and less data available causes the parameter estimates to have larger errors. Furthermore, Wheeler, Engelhard, et al. (2022a) showed the relationship between the number of responses and the average length of responses on the recovery of the topic proportions. Therefore, if the CR items on the assessment only require a few words to respond, then extracting features for each item may not be reliable. When this is the

case, the responses to each CR item in the scoring set might need to be concatenated, and a single LDA model should to extract the features for each student. Under the LDA model, it is possible that the assumption that the extracted features preserve the full information on the final response can be violated in the case that LDA does not fit well to the observed textual response.

**Issues related to fitting regression models.**

Step 1 and Step 2 of the proposed scoring method require selecting an appropriate regression method. Due to the high dimensionality of the process features, $\boldsymbol{x}_{score}$, Zhang et al. (2021) suggest using a Ridge regression with variable selection or Lasso regression where some regression coefficients are shrunk to 0. These two regression methods imply that the features and scores have a linear relationship. Previous studies, however, have suggested that the LSA features may have a nonlinear relationship with the scores (Han et al., 2020; L. Wang & Wan, 2011). Many past studies have used more robust regression techniques that allow for nonlinear relationships including support vector machines (SVMs) and support vector regressions (SVRs) with various kernels. Regarding LDA features, there are limited studies that have investigated the relationship between features and scores on CR items. Choi et al. (2017) showed that it is typical for LDA features to have low correlations with scores. The reason for the low correlation is due to the ipsative nature of the LDA features. That is, the topic proportions, which displays the proportional amount of each topic used by a response sums to one. Therefore, an examinee cannot have high proportions for more than one topic. Due to this ipsative, nonlinear relationship, CR scores are modeled using SVMs when using features from LDA. Thus, we expect the nonlinear regression methods, such as SVR with a nonlinear kernel, may work better than linear regression methods in Step 1 of the proposed scoring procedure.

**Issues related to constructing item sets.**

The last issue is about how to create the training and scoring sets. As mentioned above, splitting the items into training and scoring sets was previously chosen at random (Zhang et al., 2021). This is a possible limitation to the process-based scoring procedure since some item splits will perform better than others. Moreover, the splitting of items is more limited for mixed-format assessments, since the process features are extracted only from the CR item. Therefore, we proposed three heuristic item splitting criteria based on previous studies and theory. The item splitting criteria are based on three aspects of mixed-format assessments: the number of CR items, the average length of the textual responses for each CR item, and the IRT item information for each item.

The first criterion we consider is the number of CR items in the scoring set. Because the process features come from only the CR items, there needs to be at least one CR item in the scoring set. One CR item, however, may not be sufficient to improve $\theta$ estimate. Therefore, we suggest that at least half of the items in the scoring set be CR items. For example, if there are seven items in the scoring set, then at least four should be CR items. This criterion also limits the size of the scoring set to no more than twice the number of CR items on the assessment.

Studies have shown that the length of responses (i.e., the number of words) improve the quality of the features extracted from the text (Wheeler, Engelhard, et al., 2022a). Additionally, it is well established that length is often used in scoring algorithms for CR items. Therefore, it is appropriate to add a criterion to handle the average length of responses to the CR items in the scoring set. The second criterion establishes that the average length of responses to the CR items in the scoring set is greater than or equal to the average length of responses to all CR items.

The third criterion considers the IRT information of each item. The information of an item in an IRT framework is inversely related to the standard error of the $\theta$ estimates. The

third criterion requires that the average information of the items in the training set be greater than or equal to the average information of all items in the assessment. This criterion focuses on the training set items since the ability estimate obtained from the training set items is regressed on the features; therefore, it is important to ensure that the training set items have high quality items, compared to the rest of the items, so that the standard error around $\theta_{train}$ is reduced.

## 4.6 Empirical Case Study

In this section, we present an application of the proposed mixed-format assessment scoring procedure to empirical data. First, we present a description of the data, the mixed-format assessment, followed by the IRT model used to calibrate and score the assessment. The primary goal of this case study is to demonstrate the issues regarding the proposed mixed-format assessment scoring procedure discussed in Section 3. Therefore, the analysis strategy addresses three research questions: (Q1) does combining the text of different CR items improve the quality of the extracted features, (Q2) what is an appropriate feature extraction method for textual data to CR items and how should they be modeled to improve ability estimates, and (Q3) are there certain criteria needed to obtain better results for scoring mixed-format assessments. Specifically, we first investigate whether separate LSA or LDA models should be used for each item or if the text of all items should be combined and a single LSA or LDA model should be used to extract features (Q1). Next, we investigate which feature extraction method, LSA or LDA, and which regression modeling technique provides better results (Q2). Finally, we apply various criteria to split the items into training and scoring sets and investigate the influence of each on the scoring results (Q3). In addition, we present discussions of the implications of using the proposed scoring procedure for educators.

### 4.6.1 Data Description

The mixed-format scoring procedure is illustrated using data collected from a 13-item mixed-format assessment where eight are MC items and five are CR items. The assessment is from a previously NSF funded study and was designed to measure science inquiry knowledge for middle-grade students. The assessment contained a pre-test and post-test form which were used to assess a learning intervention. The original study had responses from more than 5,000 examinees. A common concern with mixed-format assessments is whether the items measure the same construct (S. Kim & Kolen, 2006). Traub (1993) suggested that if the CR items are evaluated on a content basis rather than on the basis of writing conventions (e.g., spelling, grammar, style), then it should be possible to determine if the MC and CR items measure the same construct. For this study, therefore, the CR items were scored based on content and did not consider writing conventions. Furthermore, results from a factor analysis found that all 13 items loaded on a single factor with eigenvalues greater than 1 and loaded on the remaining factors with eigenvalues less than 1. These findings support us to calibrate the item parameters for 13 items using unidimensional IRT models.

For this study, all 5,000 examinees and their responses to the pre-test were used to calibrate the item parameters for the 13 items. The eight MC items and two CR items were scored dichotomously. For these items, a two-parameter logistic IRT model (2PL; Birnbaum, 1968; Lord & Novick, 1968) was used to estimate item parameters (see Appendix B). The remaining three CR items were scored using three rating categories. For these items, a graded-response model (GRM; Samejima, 1969, 1997) was used to estimate the item parameters. All five CR items were scored by raters using a rubric. The raters went through a rigorous rater-training protocol to reduce the amount of rater bias in the scores. The five CR items received a 94-percent agreement among the raters. Table 4.1 shows the 13 items on the assessment, their item types, the number of scoring categories, and the IRT model

Table 4.1: Description of items on the assessment for the empirical example

| Item Number | Item Type | Number of Categories | IRT Model |
|---|---|---|---|
| 1 | MC Item | 2 | 2PL |
| 2 | MC Item | 2 | 2PL |
| 3 | MC Item | 2 | 2PL |
| 4 | MC Item | 2 | 2PL |
| 5 | MC Item | 2 | 2PL |
| 6 | CR Item | 2 | 2PL |
| 7 | CR Item | 3 | GRM |
| 8 | MC Item | 2 | 2PL |
| 9 | MC Item | 2 | 2PL |
| 10 | CR Item | 3 | GRM |
| 11 | MC Item | 2 | 2PL |
| 12 | CR Item | 3 | GRM |
| 13 | CR Item | 2 | 2PL |

used to estimate the item parameters.

A subset of 530 examinees' responses were used from the initial 5,000 examinees to investigate our research questions. The 530 examinees were selected by eliminating students that had responses containing missing data. Additionally, the textual data for all examinees in this subset had written responses to the CR items that allowed for features to be extracted. The features were extracted from the textual data with two different methods, the LSA and LDA methods. In addition, the scored responses to all items on the assessment were used to obtain ability estimates using both the traditional IRT scoring procedure using EAP and the proposed mixed-format assessment procedure.

### 4.6.2 Item Response Models

For this study, two item response models were employed. A 2PL was used to score the dichotomous items and a graded response model GRM was used to score the polytomous

items. The item parameters for the 2PL and GRM were estimated using flexMirt (Cai, 2013), which implements an expectation-maximization (EM) algorithm with fixed quadratures.

### Two Parameter Logistic Model

For the dichotomously scored items, the two response categories were defined as an incorrect or correct response. The 2PL model can be described using a logistic form to model the probability of a correct response as a function of the examinee's ability, item discrimination, and item difficulty. The item discrimination is a proxy for the correlation between the item and the underlying construct, and item difficulty is the location of the item on the underlying construct. The item response function for the 2PL is expressed as

$$P(Y_j = 1|\theta) = \frac{1}{1 + \exp\{a_j(\theta - b_j)\}}, \tag{4.10}$$

where $P(Y_j = 1|\theta)$ is the probability of a correct response on item $j$ for an ability of $\theta$; $a_j$ is the item discrimination parameter for item $j$; and $b_j$ is the item difficulty parameter for item $j$.

### Graded Response Model

The GRM is used for polytomous items where the response categories are ordered. The GRM assumes that an item has $m_j$ ordered response categories and an examinee is only allowed to select one of the categories. For CR items, a category is typically assigned to an examinee through a rater or automated scoring engine where lower scores are associated to a more incorrect response and higher scores are associated to a more correct response. Samejima (1997) described the GRM with a logistic form using the boundary characteristic curve (BCC). The BCC for an $m_j$ ordered response item is expressed using the following function:

$$P_{jk}(Y = k|\theta) = \frac{1}{1 + \exp\{-a_j(\theta - b_{jk})\}}, \tag{4.11}$$

where $P_{jk}(Y = k|\theta)$ is the probability of the response to item $j$ for category $k$, $k = 1, ..., m_j$; $\theta$ is the ability of the examinee; $a_j$ is the item discrimination parameter for item $i$; and $b_{jk}$ is the location of category $k$ for item $j$ on the underlying construct. As the ability of an examinee increases, the probability of being assigned a higher category also increases.

### 4.6.3 Selecting the number of features for LSA and LDA.

For LSA, the number of features for each item is determined based on a proportion of the total sum of the singular values of the selected features compared to the total sum of all singular values. This shows the amount of variation explained by the selected features. Specifically, the number of features whose singular values had a total summed proportion of .4, compared to the total sum of singular values, were extracted. This selection criterion is often used for LSA models when they are used as the basis of rubric scoring CR items (e.g., Evangelopoulos et al., 2012).

For the LDA feature extraction, the number of features for each item is determined by the deviance information criterion (DIC; Spiegelhalter et al., 2002). Specifically, in this example, a two-topic to ten-topic LDA model is estimated and the number of topics with the smallest DIC value is selected. This selection criteria is often used for LDA models in educational measurement (e.g., Cardozo-Gaibisso et al., 2019; Wheeler, Engelhard, et al., 2022a).

### 4.6.4 Estimating Ability using Proposed Scoring Procedure

The empirical example is meant to illustrate how the new scoring procedure for mixed-format assessments could be used in practice. Once the items parameters are estimated using the appropriate IRT models, they are held fixed for the scoring procedure, similar to a calibrated item bank. Since the true ability for each examinee is unknown, five of the thirteen items (Items 1, 4, 5, 8, and 11) are used to calculate a reference ability estimate for each of the 530 examinees. These are denoted as $\hat{\theta}_{reference}$. These ability estimates are treated as a reference

point when investigating the effectiveness of the new scoring procedure relative to standard IRT scoring. The five reference items were randomly selected after controlling for their item information, which is used as a proxy for the quality of the items. That is, the average item information of the five reference items was constrained to be equal to, or greater than, the average item information of the remaining items. The remaining eight items, consisting of three MC items and five CR items, were randomly split into the training and scoring sets. The process features were extracted separately from the CR items in the scoring set using LSA and LDA. The process-based ability parameters were estimated for each examinee using the proposed scoring procedure described in Section 3.1.

In practice, the items would be randomly split into the training and scoring set, however, for the purpose of investigating the research questions in this study, all possible combinations of training and scoring sets were used with the constraint that at least one CR item be included in the scoring set. Table 4.2 shows the total number of unique item sets for the different sizes of training and scoring sets.

Table 4.2: Number of different training and scoring item sets

| Number of Items (Training Set) | Number of Items (Scoring Set) | Number of combinations |
|---|---|---|
| 7 | 1 | 5 |
| 6 | 2 | 25 |
| 5 | 3 | 55 |
| 4 | 4 | 70 |
| 3 | 5 | 56 |
| 2 | 6 | 28 |
| 1 | 7 | 8 |

*Note.* The training and scoring sets have to have at least one item; the scoring set has to have at least one CR item; there were a total of 247 combinations to divide the eight items into training and scoring sets

As can be seen in Table 4.2 there are seven different sizes of training and scoring sets. The number of combinations for each size range from 5 to 70. For example, when there are seven items in the training set and one item in the scoring set, then there are only 5 possible

combinations in which the scoring set has at least one CR item, namely, the five CR items. There are a total of 247 combinations to split the eight items into a training and scoring set where the scoring set has at least one CR item. Each research question was investigated using all 247 combinations, meaning 247 different process-based ability parameters ($\hat{\theta}_{process}$) were estimated for each examinee using the proposed scoring procedure.

### 4.6.5 Evaluation Technique

The evaluation process for each research question follows the technique used by Zhang et al. (2021) which relies on comparing the estimates obtained from standard IRT scoring and process-based scoring to the reference ability estimates. For each of the 247 combinations, a process-based ability estimate is obtained using the proposed scoring procedure, denoted as $\hat{\theta}_{process}$. Additionally, the scoring set items for each combination are used to obtain an ability estimate for each examinee using standard IRT scoring, denoted as $\hat{\theta}_{IRT}$. The ability estimates obtained from the proposed scoring procedure and standard IRT scoring are compared to the reference ability estimate using mean squared error (MSE). The MSE for the proposed scoring procedure and standard IRT scoring are defined as

$$MSE_{process} = \frac{\sum_{i=1}^{N}(\hat{\theta}_{i,process} - \hat{\theta}_{i,reference})^2}{N} \tag{4.12}$$

$$MSE_{IRT} = \frac{\sum_{i=1}^{N}(\hat{\theta}_{i,IRT} - \hat{\theta}_{i,reference})^2}{N}, \tag{4.13}$$

where $MSE_{process}$ is the MSE between the ability estimates obtained from the process scoring procedure and the reference ability estimates; and $MSE_{IRT}$ is the MSE between the ability estimates obtained from standard IRT scoring and the reference ability estimates. The MSEs between the two scoring procedures are then compared to each other, where a lower MSE indicates the ability estimates are more accurate with respect to the reference ability estimate.

In addition to the MSEs, we compared the ability estimates between all three scoring methods using the Pearson correlation coefficient. Specifically, for every combination of item sets, we found the correlation between the ability estimates obtained from standard IRT scoring, the proposed scoring procedure with LSA, and the proposed scoring procedure with LDA. A high Pearson correlation between two ability estimation results implies that they are high linearly correlated with each other.

### 4.6.6 Results

#### Descriptive Analysis of Items

Before investigating the research questions outlined at the beginning of the empirical study, a descriptive analysis was conducted to provide information about the five CR items on the assessment. Table 4.3 shows various descriptive statistics about each of the items on the assessment: the average response length to each CR item, which is calculated after removing stop words and stemming; the number of extracted features from the LSA model, which is determined based on the criterion that the singular values accounted for .4 of the variation; the average absolute correlation between the LSA features and the total score; the number of features extracted from the LDA model, informed by DIC; the average absolute correlation between the LDA features and the total score.

Table 4.3: Descriptive statistics of CR items

| | | LSA | | LDA | |
|---|---|---|---|---|---|
| Item | Average Response Length | Number of Features | Average Feature Correlation | Number of Features | Average Feature Correlation |
| 6 | 10.61 (5.12) | 17 | 0.15 (0.05) | 5 | 0.07 (0.03) |
| 7 | 4.17 (1.02) | 15 | 0.11 (0.03) | 5 | 0.03 (0.03) |
| 10 | 3.78 (1.33) | 15 | 0.11 (0.02) | 5 | 0.04 (0.02) |
| 12 | 4.31 (1.68) | 13 | 0.09 (0.02) | 5 | 0.10 (0.04) |
| 13 | 5.18 (1.87) | 16 | 0.12 (0.03) | 5 | 0.13 (0.03) |

*Note.* The average response length is determined post data cleaning and text processing. The average feature correlation is based on the absolute values of the correlation of each feature with the total score.

Investigating the average absolute correlation between the LSA and LDA features and the total score is an important initial step for this scoring procedures, as this will help us

identify an appropriate model for the first regression step (Step 1) outlined in Section 3.1. The average correlations shown in Table 4.3 are small, although this is often the case for features extracted using LSA and LDA (Choi et al., 2017; Xiong et al., 2021). For this empirical example, both ridge regression and SVR were utilized to regress $\hat{\theta}_{train}$ on $\boldsymbol{x}_{score}$. The reason to select ridge regression is because text features are chosen through a model selection process and the LSA extracted a large number of features, thus in this context ridge regression can function as a variable selection approach. The ridge regression also represents an example of modeling the features and scores via a linear relationship. The SVR was used because support vectors are common techniques used by machine scoring algorithms for text classification (H. Kim et al., 2005). SVR represents an example of modeling the features and scores via a nonlinear relationship.

**Research Question One: Combining Responses to CR Items**

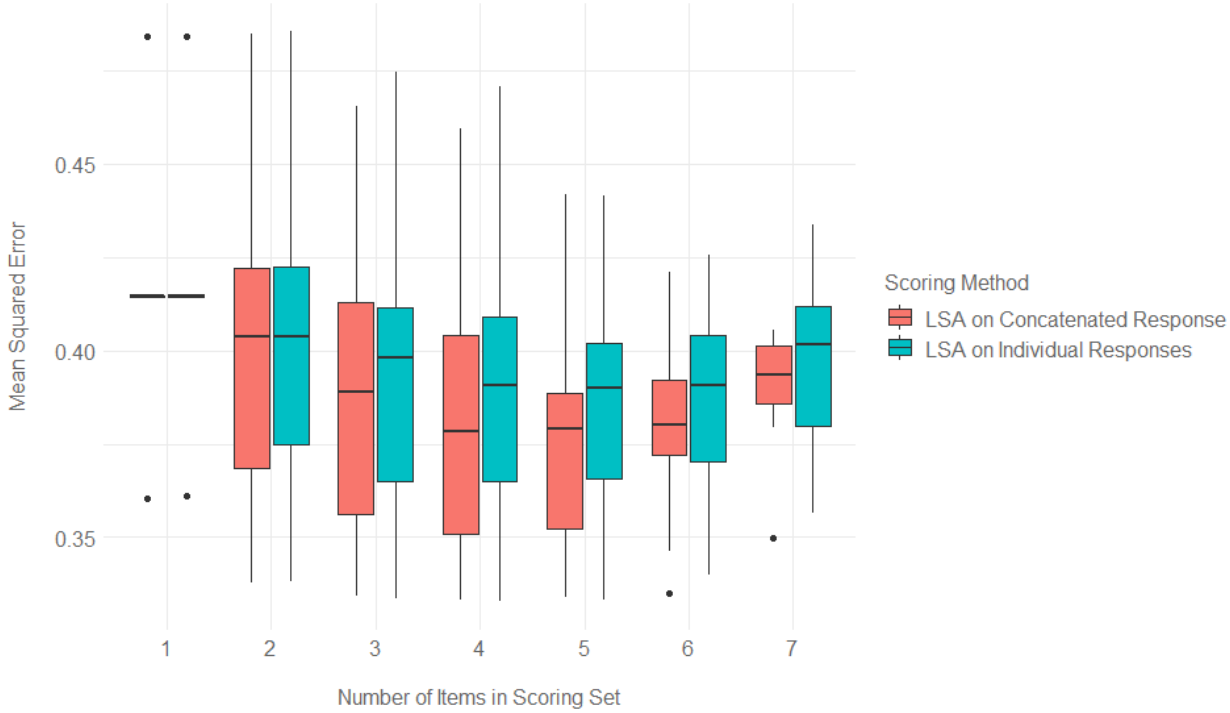
The investigation of this first question analyzed the performance of the scoring procedures, when features were extracted for each CR item separately, and when features were extracted by combining the CR items. When features were extracted for each CR item, a separate LSA and a separate LDA model were estimated from the textual responses to each CR item. A set of features were then extracted for each item and then the features for each CR item in the scoring set were augmented to create a single set of features. The latter step combines the responses to the CR items in the scoring set and fits a single LSA and a single LDA model to the combined responses. The features extracted from the combined responses were used as the set of features in the scoring procedure.

The performance of the mixed-format scoring procedure, in which features are extracted from the responses to each CR item in the scoring set and then augmented into a single vector of process features, is compared to the performance of the scoring procedure in which features are extracted from all CR items in the scoring set as a single combined response. The

MSEs for all possible training and scoring set combinations are calculated for the proposed scoring procedures with both of the methods for treating the textual responses.

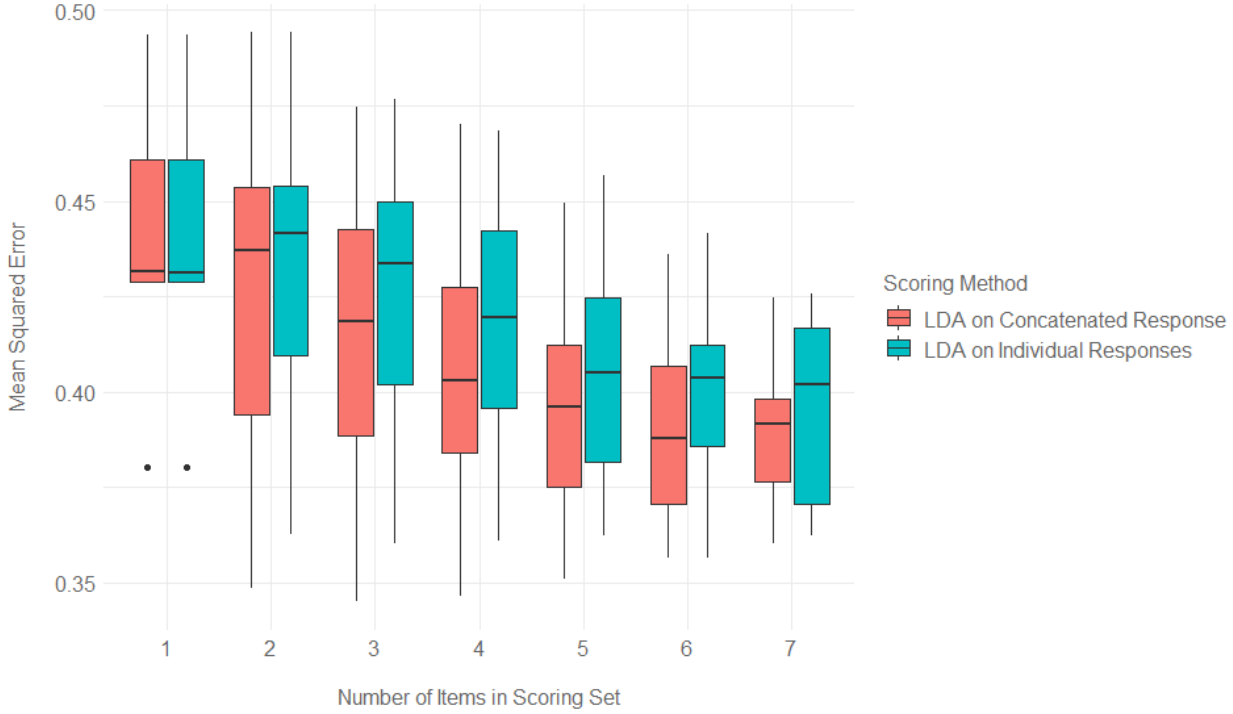
Figure 4.2 displays a series of grouped box plots for the MSEs of the different scoring sets. The MSEs are calculated using Equation (4.12) where the process-based ability estimates are obtained from the proposed scoring procedure using the LSA feature extraction method. For each scoring set item size there are two box plots, one for the proposed scoring procedure where separate features are extracted from the responses to each CR item (denoted as *LSA on Individual Responses* in Figure 4.2), and the other is for the proposed scoring procedure where a single set of features are extracted from the combined responses to the CR items in the scoring set (denoted as *LSA on Combined Response* in Figure 4.2). The number of measures in each box plot is the number of possible combinations for the given number of items in scoring set which is reported in Table 4.2.

Figure 4.2: MSEs for features extracted from each CR item vs. features extracted from the combined response using LSA



The box plots displayed in Figure 4.2 show that there is little difference in MSEs between the two methods for treating the textual data. That is, for LSA, it does not seem to matter if the features are extracted from each individual CR item and then combined or if the features are extracted from a single LSA model where the responses to all CR items in the scoring set are combined. Similarly, Figure 4.3 also displays a series of box plots for the MSEs of the different scoring sets. The MSEs are calculated using Equation (4.12) where the process-based ability estimates are obtained using features extracted from LDA. Similar to the results for LSA, the box plots displayed in Figure 4.3 show that there is little difference in MSEs between the two methods for treating the textual data for LDA.

Figure 4.3: MSEs for features extracted from each CR item vs. features extracted from combined response using LDA




The subsequent results presented use the combined responses due to the computational savings for fitting a single topic model over fitting multiple topic models for each CR item,

and there does not appear to be a significant difference in the results between the two ways of fitting the topic models.

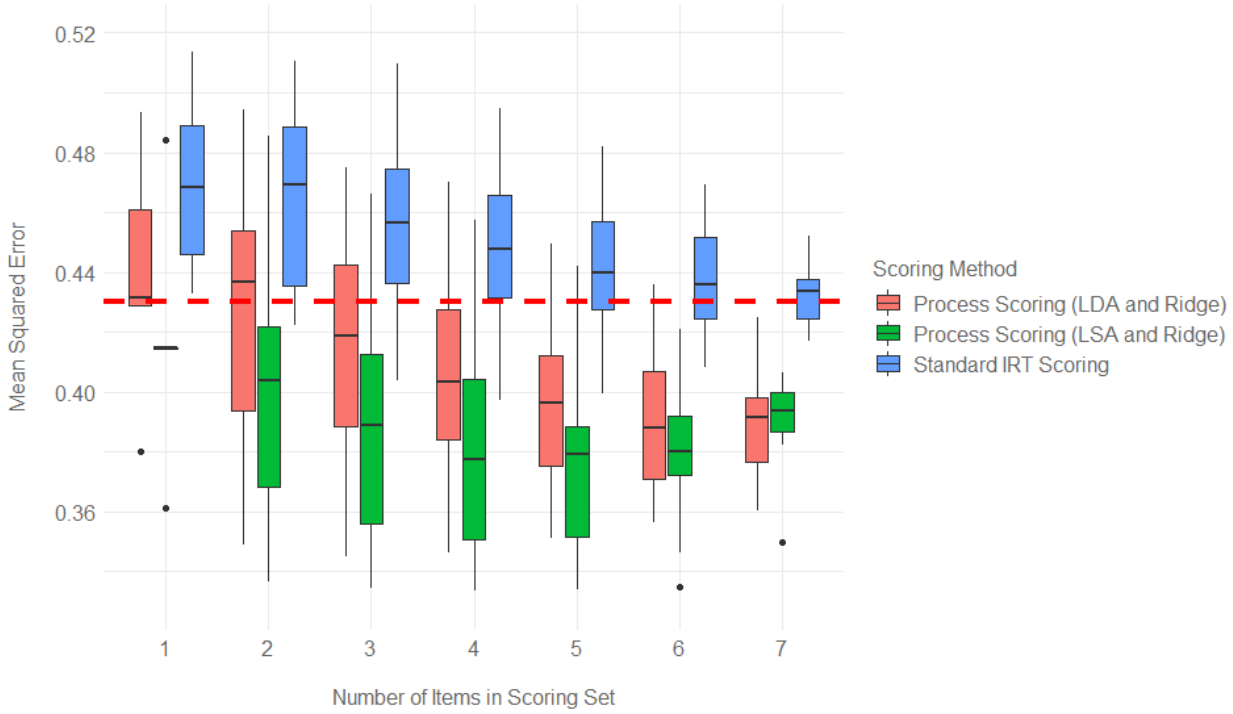
### Research Question Two: Latent Feature Extraction Methods for CR Items

The second research question investigates the performance of the proposed scoring procedure using the two feature extraction methods for CR items. The performance of both methods are compared to each other and also to a standard IRT scoring procedure using EAP. Two process-based ability estimates were obtained for each examinee from the proposed scoring procedure using the two different feature extraction methods. The two process-based ability estimates were compared to the reference ability estimates using MSE following Equation (4.12). Additionally, ability estimates for each examinee were obtained from a standard IRT scoring procedure using the items in the scoring set. These are compared to the reference estimates using MSE following Equation (4.13). The MSEs across the different scoring approaches are compared in a series of grouped box plots. In addition, the use of different regression techniques is also compared. Specifically, both ridge regression and SVR were used for Step 1 of the proposed scoring procedure (refer to Section 3.1).

First, ability estimates were obtained using the proposed scoring procedure with each of the different feature extraction methods and ridge regression. Figure 4.4 displays a series of grouped box plots for the MSEs obtained from the scoring procedure using LSA and ridge regression, the proposed scoring procedure using LDA and ridge regression, and a standard IRT scoring procedure. The red dashed line represents the MSE between the reference abilities and the ability estimates obtained from a standard IRT scoring procedure using all eight items.

The results in Figure 4.4 show that the proposed scoring procedure using the LSA and ridge regression generally yielded a more accurate ability estimates as reflected in the lower MSEs than either the proposed process scoring using LDA or the standard IRT scoring

Figure 4.4: MSEs for the proposed scoring procedure using LSA and ridge regression, LDA and ridge regression, and standard IRT scoring

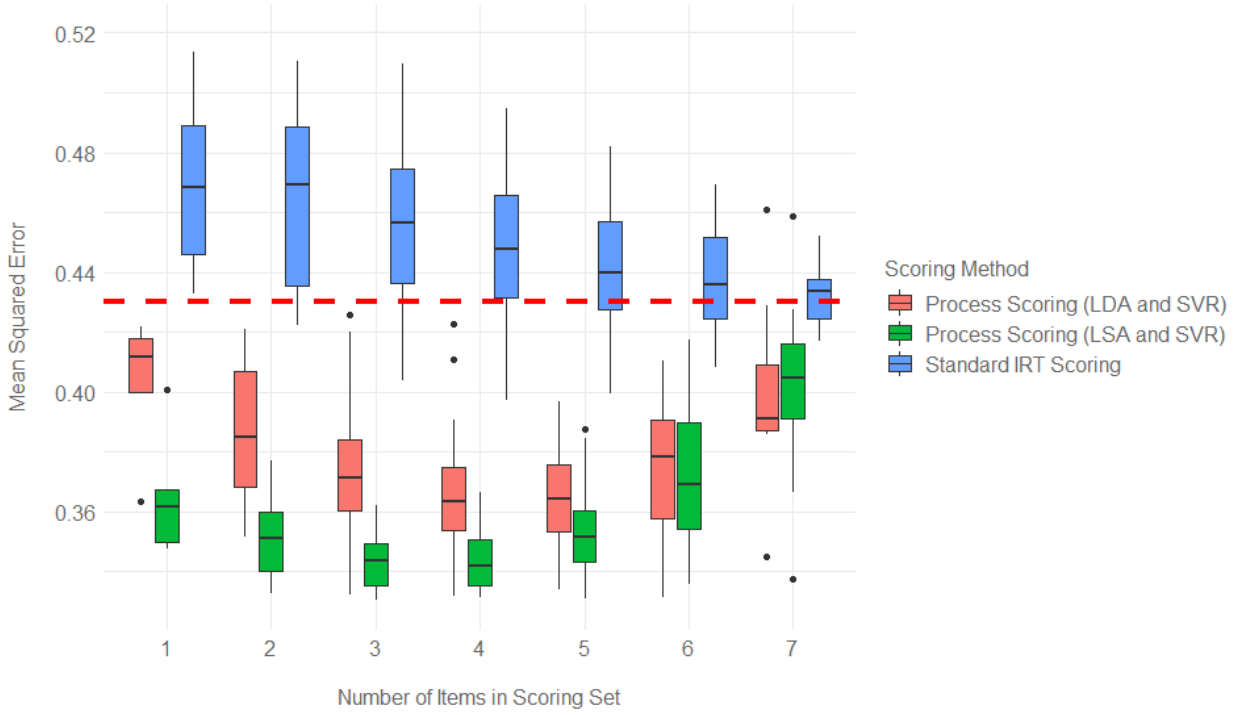



method. The results also show that, regardless of the number of items in the scoring set, the proposed scoring procedure using LSA and ridge regression obtained better ability estimates, on average, than the standard IRT scoring on all eight items (as reflected in the LSA box plots below the red-dashed line in Figure 4.4). Additionally, the results suggest that the LSA feature extraction method generally provided better ability estimates than the LDA feature extraction method for the proposed scoring procedure.

Second, ability estimates were obtained using the proposed scoring procedure with the different feature extraction methods and SVR. Figure 4.5 displays a series of grouped box plots for the MSEs obtained from the scoring procedure using LSA and SVR, the proposed scoring procedure using LDA and SVR, and a standard IRT scoring procedure. The red

dashed line represents the MSE between the reference abilities and the ability estimates obtained from a standard IRT scoring procedure using all eight items.

Figure 4.5: MSEs for the proposed scoring procedure using LSA and SVR, LDA and SVR, and standard IRT scoring




Similar to the ridge regression results, the proposed scoring procedure using LSA generally yielded more accurate ability estimates than the proposed scoring procedure using LDA or standard IRT scoring. In addition, the LDA based process scoring results with SVR also consistently had a better ability estimation than the standard IRT scoring, which were not observed in the ridge regression based results. Finally, the results from the SVR approach seemed to yield lower MSE values than the results from the ridge regression approach, regardless of the feature extraction method. This suggests that perhaps the features extracted using LSA and LDA are not linearly related to the ability estimates, which is consistent with previous findings (Ramesh & Sanampudi, 2021).

One important observation from these results, particularly evident in Figure 4.5, is that there seems to be a slight increasing trend to the average MSEs for the proposed scoring procedure as the number of items in the scoring set increases. The reason for this increase in MSE as the size of scoring set increases is due to the decrease of items in the training set. Recall from Section 3.1, the first regression step in the proposed scoring procedure is to regress the ability estimates obtained from the training set items on the features extracted from the CR items in the scoring set. Thus, as the number of items in the training set decreases so does the possible outcome space of the ability estimates using the training set items. Further, as the possible outcomes for ability estimates decrease for the training set items, so does the variability in the dependent variable for the first regression step of the scoring procedure. In addition, the lower number of items in the training set also means that there is more uncertainty around our ability estimates obtained from those items, which causes the results to be less accurate from the proposed scoring procedure.

**Research Question Three: Applying Criteria to Split Items**

The three item splitting criteria, described in Section 4.3, were investigated using an approach similar to that used for the previous research questions. The proposed scoring procedure using LSA and LDA as feature extraction methods and using ridge regression and SVR as modeling techniques was applied to all combinations of training and scoring sets. Then the item splitting criteria were applied sequentially with the the number of CR items in the scoring set first, the criterion for average length of responses to the CR items next, and the criterion for the item information last. The combinations of training and scoring sets that did not satisfy the criteria were removed from the results.

Figure 4.6 displays similar grouped box plots to those presented above, however, separate plots are displayed to show the results of sequentially applying each criterion. The first quadrant of the figure, labeled *No Criteria*, are the same MSEs as shown above for the

proposed scoring procedure with LSA and ridge, the proposed scoring procedure with LDA and ridge, and a standard IRT scoring procedure. The second quadrant, labeled *First Criterion*, are the remaining combinations of the training and scoring sets after applying the number of CR items criterion. The third quadrant, labeled *First and Second Criteria*, are the remaining combinations of the training and scoring sets after applying the number of CR items criterion and the average length of responses criterion. The fourth quadrant, labeled *All Criteria*, are the remaining combinations of the training and scoring sets after applying the number of CR items criterion, the average length of responses criterion, and the item information criterion. The four plots in Figure 4.6 suggest that sequentially applying the item splitting criteria may help improve the performance of the proposed scoring procedure when using ridge regression. Specifically, the average response length criterion and the item information criterion decreased the variance in MSEs for the proposed scoring procedure.

Figure 4.6: Results for applying item splitting criteria for scoring procedure using ridge regression

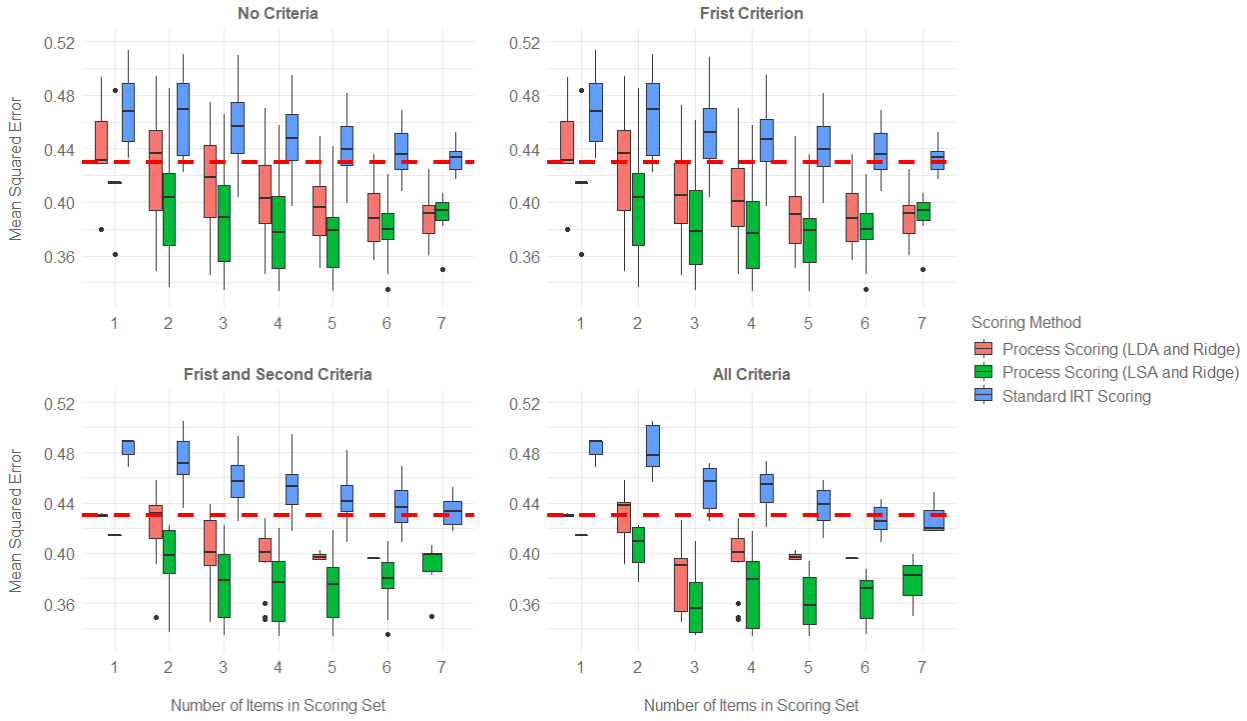
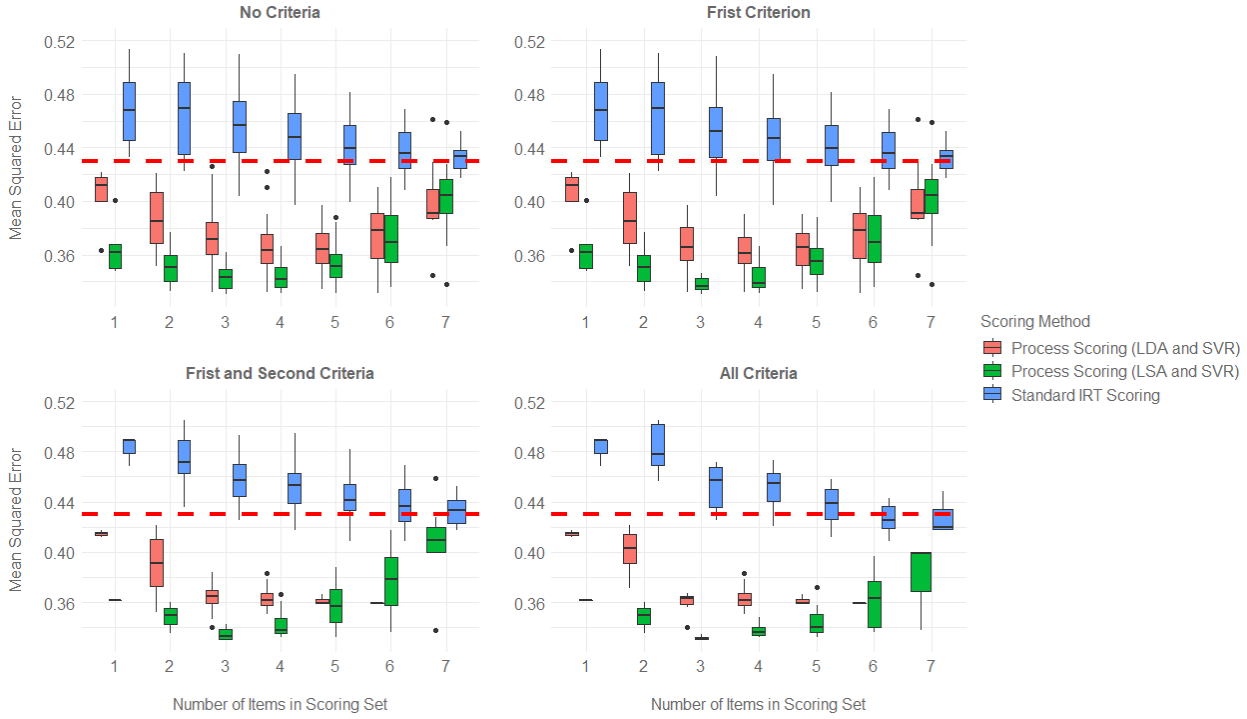


Figure 4.7 shows the change in MSE for sequentially applying the three criteria to the proposed scoring procedure using SVR. The quadrants in Figure 4.7 are the same as described above. Results and trends from this analysis are similar to the ones found in Figure 4.6, meaning that sequentially applying the proposed criteria may help improve the performance of the proposed scoring procedure, regardless of the regression method used.

Figure 4.7: Results for applying item splitting criteria for scoring procedure using SVR




## 4.6.7 Practical Implications of Results

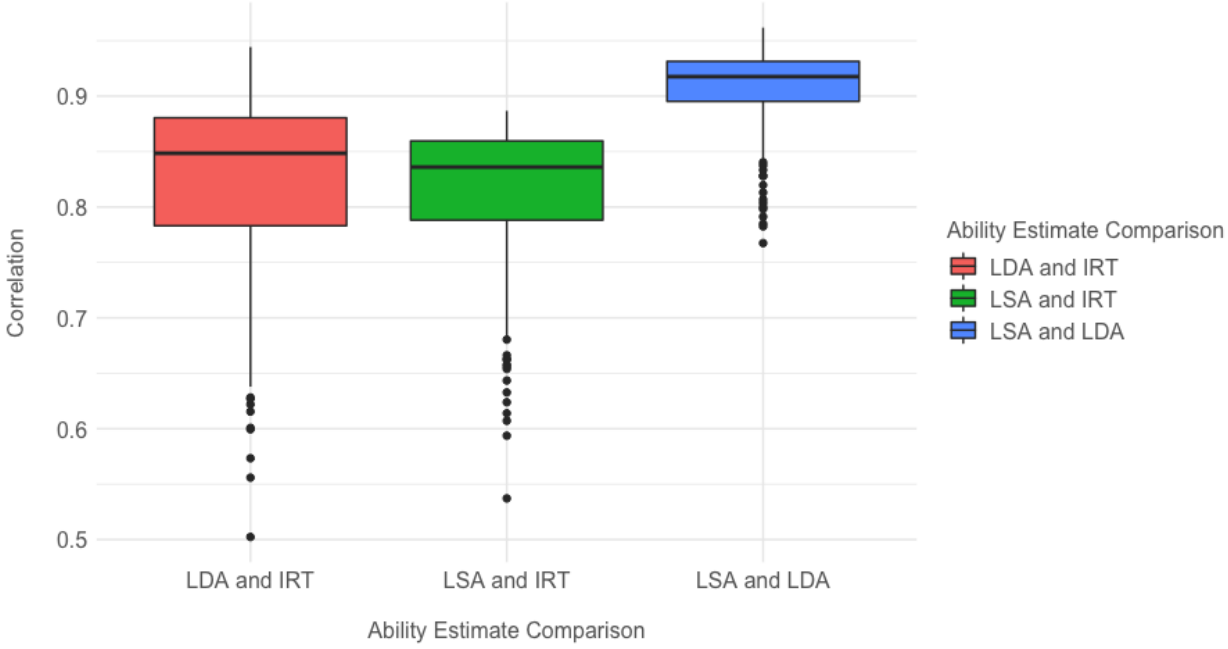
In addition to demonstrate that the proposed new scoring for mixed-format assessments improves the ability estimates for examinees, we provide a detailed discussion regarding the implications of using the new scoring procedure in practice in this section.

We first investigate the differences between the LSA and LDA feature extraction methods and their impacts on the ability estimates. Previous research showed that the latent semantic

spaces between LSA and LDA are highly similar (Wheeler, Cohen, et al., 2021). Results from this study showed that the LSA method performs better than LDA, however, both feature extraction methods improved ability estimates over standard IRT scoring.

Calculating the Pearson correlation between the ability estimates obtained from each scoring method provides a better understanding into the relationship between the proposed scoring procedure and standard IRT scoring. Figure 4.8 displays three boxplots that contain the correlation between ability estimates for all combinations of item sets in this study. The first boxplot shows the correlation between the proposed scoring procedure using LSA and standard IRT, the second boxplot shows the correlation between the proposed scoring procedure using LDA and standard IRT, and the third boxplot shows the correlation between the proposed scoring procedure using LSA and LDA.

Figure 4.8: Correlation between ability estimates obtained for each scoring method




Although the correlations between the three scoring methods are relatively high, the correlation between the proposed scoring procedure using LSA and LDA are higher than the correlations between the proposed scoring procedure and IRT scoring. This suggests

that both feature extraction methods yield relatively similar ability estimates, which can also be reflected by the similar MSE values from results shown in Figure 4.4 and Figure 4.5. Additionally, the variance of the correlations is smaller suggesting that the ability estimates from the proposed scoring procedure are more consistent.

Since LSA and LDA provide similar results, we used the LSA results to demonstrate the practical utility of the proposed scoring procedure. We investigated how the proposed scoring procedure can be used in practice by looking at a specific example from the empirical study. Table 4.4 shows three examinees with the same response patterns. Listed in the table are the ability estimate from a standard IRT scoring procedure, the ability estimate from the proposed scoring procedure using LSA and ridge regression, and the reference ability estimate.

Table 4.4: Sample of three examinees response data from the empirical example

| Examinee | Scoring Set Items | Train Set Items | Response Patterns (Scoring Set) | Response Patterns (Train Set) | Standard IRT Ability | Process-Based (LSA and Ridge) Ability | Reference Ability |
|---|---|---|---|---|---|---|---|
| A | 6, 9, 10, 12 | 2, 3, 7, 13 | 0000 | 1000 | -1.07 | -1.37 | -1.54 |
| B | 6, 9, 10, 12 | 2, 3, 7, 13 | 0000 | 1000 | -1.07 | -0.48 | -0.55 |
| C | 6, 9, 10, 12 | 2, 3, 7, 13 | 0000 | 1000 | -1.07 | -1.05 | -0.93 |

Under standard IRT scoring, these three examinees would have the same ability estimates, however, under the new scoring procedure they have quite different ability estimates. The ability estimate for Examinee A under the proposed scoring procedure decreased compared to that from the standard IRT scoring ($\hat{\theta}_{IRT} = -1.07$ and $\hat{\theta}_{process} = -1.37$). The ability estimate for Examinee B increased under the proposed scoring procedure ($\hat{\theta}_{IRT} = -1.07$ and $\hat{\theta}_{process} = -0.48$). Examinee C had similar estimates for the proposed scoring procedure and traditional IRT scoring procedure ($\hat{\theta}_{IRT} = -1.07$ and $\hat{\theta}_{process} = -1.05$).

The difference in ability estimates between the proposed scoring procedure and standard IRT scoring is due to the features extracted from the responses of the examinees. That is, the textual responses provide additional information that can be obtained through various feature extraction methods and used to help infer the latent ability of students. In fact,

the ability estimates from the proposed scoring procedure are closer to the reference ability. Furthermore, reading the textual responses from these three students can help determine the difference between their extracted features. Table 4.5 shows the textual responses of the three examinees to the CR items in the scoring set. All three students responded to Item 6 but not to Items 10 and 12; and all three students incorrectly answered Item 6, but for different reasons. Figure 4.9 shows the operational form of Item 6 that students encountered.

Table 4.5: Sample of textual responses to the CR items in the scoring set

| Student | Item 6 | Item 10 | Item 12 |
|---|---|---|---|
| A | the effect on the rubberband will be that he is using too much kinetic energy on it also he have to use less energy on the rubberband when he pulls it may go to far. | BLANK | BLANK |
| B | The effect was that when Jorge releases the rubberband the potential energy increase it is being stored in the slingshot and the stone until it is release. | BLANK | BLANK |
| C | the will travel but not that far | BLANK | BLANK |

Item 6 on the science inquiry assessments prompts students to write about their understanding of the energy effects of a slingshot. Based on the rubric, students are assessed on the science inquiry domain and are dichotomously scored for Item 6 (either correct or incorrect). Examinees received a correct score from raters, if they correctly identify potential and kinetic energy, and they are able to connect those two concepts together within a written slingshot experiment. The response of Examinee A is incorrect because they were unable to identify the effects of potential and kinetic energy and their connection. In fact, their response indicates that they have a misunderstanding of kinetic energy. The proposed scoring procedure was able to identify this through the extracted features from the text and their ability was lowered. The response of Student B is incorrect because they confused potential energy with the release of the slingshot and not with the pulling back of the slingshot, however, their

Figure 4.9: Operational form of Item 6 on the science inquiry assessment

**Slingshot Experiment**

Jorge tells his friend Brian that he is confused about the difference between potential energy (stored energy) and kinetic energy (energy in motion). Brian gets his slingshot and says that they can use the slingshot and a stone to understand the difference between these two forms of energy. The two boys practice putting a stone in the slingshot, pulling back the rubber band, and releasing the rubber band to watch the stone fly through the air.

Choose the best answer for question 9 and write your answer on the answer sheet.

9) What is the ***effect*** on the slingshot and stone when Jorge pulls back on the rubber band?

A) Potential energy decreases because Jorge is using up energy to pull the rubber band.
B) Potential energy increases because energy is being stored in the slingshot and the stone.
C) Kinetic energy increases because it is stored in the slingshot and the stone until it is released.
D) Kinetic energy is not involved because there is no motor in the slingshot and the stone.

Next, write your answer for question 10 in the box on the answer sheet.

10) What is the ***effect*** on the slingshot and stone when Jorge releases the rubber band? (Include potential or kinetic energy in your description.)

response indicates that they understand the concept of potential energy being stored in the slingshot and the connection between potential and kinetic energy. Again, the proposed scoring procedure was able to identify key concepts in the writing through the extracted features and provide Student B with an improved ability estimate. Lastly, the response of Student C is incorrect because they did not answer the question directly and used "everyday" language (i.e., nonacademic/nonscientific language), however, it is unclear whether or not Student C has any misunderstanding about potential and kinetic energy. This is identified

by the proposed scoring procedure and the ability estimate did not differ much from the estimated provided under standard IRT scoring.

Although the proposed scoring procedure estimated different abilities for these three students, who would otherwise have the same ability estimate under standard IRT, it is clear from their responses that they all possess a slightly different understanding of science inquiry. For example, based on the responses from Student A and Student B, it can be seen that Student B has a better understand of science inquiry than Student A. Perhaps, the proposed scoring procedure was able to identify the differences between these two students and was able to estimate better ability estimates. An important benefit of the proposed scoring procedure is that it is able to identify differences in the responses to CR items and incorporate them into the estimation of ability. Under traditional IRT scoring, a wrong answer is counted the same regardless of the reasons why the answer was wrong, however, it the proposed scoring procedure is able to pick up on the various, and perhaps subtle, nuances of the responses to each CR item and adjust the ability estimates of each student accordingly.

## 4.7 Discussion

Our study proposed a framework for scoring mixed-format assessments that combines additional information from the textual responses to CR items, which are typically left unanalyzed once the item is scored, with scores and ability obtained from traditional IRT models. In this setup, features are extracted from each essay and used in a two-step regression procedure with the IRT ability estimates as the dependent variable. Through an empirical example, we investigated three research questions. The results from which demonstrate how the proposed scoring procedure can be used for mixed-format assessments and suggest that additional information can be extracted from the responses in order to provide more accurate estimates for the ability of examinees. The conclusions draw upon the three questions can also provide

educators and practitioners with useful guidelines for applying our proposed process-based scoring procedure to mixed-format assessments.

The first research question we investigated is regarding whether separate LSA or LDA models should be used for each item or if the text of all items should be combined and a single LSA or LDA model should be used to extract. The results for both LSA and LDA in our empirical example show that there is little difference in performance of the proposed scoring procedure between extracting a set of features for each CR item and extracting a single set of features for the concatenated responses for both LSA and LDA. However, we want to point out that this finding may depend on the amount of words that appear in a single CR answer. When a CR item only requires a few number of words, then more responses (i.e., larger sample size) are needed to get stable estimates of the latent structure of the text, however, if the all CR items are treated as a single response then the number of words for a response increases which provides a more stable estimate of the latent structure of the text. Furthermore, the features for the LDA model become unstable and have large standard errors, if there are not enough data to estimate the parameters of the assumed model (Wheeler, Cohen, et al., 2021). For these reasons, we recommend to treat all CR items as a single response when extracting features, rather than extracting features for each constructed response item.

The second research question is on comparing the proposed scoring procedure using features expected by LSA and LDA, and investigating two regression procedures, ridge and SVR. Results from our empirical analysis indicates that both LSA and LDA are appropriate to be used for the proposed process scoring procedure, and the LSA features based results have better performance than the LDA based results in different conditions. In terms of the regression method that models how the features are related to ability estimates, the SVR method has a better result than the ridge regression method in our set up, which is more obvious for the LDA feature results. This might imply a non-linear relationship between the textual features and final scores, especially for the features extracted by LDA. Based on

these findings, we suggest to use both LSA and LDA but combine with SVR approach in the proposed process scoring procedure for mixed format assessment.

The third research question explores how to construct the scoring and training sets for the proposed process scoring procedure for mixed format assessment. In an operational testing setting, randomly selecting the training and scoring item sets may not be optimal due to the relatively large variance of MSEs observed in the previous results. Therefore, we proposed three heuristic item splitting criteria that related to (1) the number of CR items in the scoring set, (2) the average length of responses to CR items, and (3) the item information. The results of applying the item splitting criteria to the proposed scoring procedure for ridge regression and SVR are similar. That is, applying the three criteria seemed to provide some guidance in how the items should be split into training and scoring sets. Application of the three criteria also appeared to increase the separation between the MSE box plots of the proposed scoring procedure and the MSE box plots of the standard IRT scoring. This suggests that the three criteria may help select training and scoring item sets that will provide the most accurate results under the proposed scoring procedure.

The general framework provided by Zhang et al. (2021) considered features extracted from data collected by the computer during the assessment process, such as log-file data, response time, and eye-tracking data. Collecting complicated log-file data can be expensive and require a more robust assessment system. One benefit, therefore, of the proposed scoring procedure is that it can be applied to paper-pencil assessments and less robust online assessment systems since the textual responses are recorded as answers. It is also important to note that this process data approach to improve ability estimates only works when the CR items are measuring the same latent ability, that is, if there are two separate latent abilities being measured by differed CR items, then this approach does not guarantee improved estimates. Though promising in using the mixed format assessment, the current study can be improved for the following directions. First, the study focused on text extraction methods, LSA and

LDA, which belong to a family of models known as topic models. Future studies can used other models from the family of topic models, supervised latent Dirichlet allocation, in order to extract process features from textual responses to improve ability estimates. In addition, for the MC items in the current study, we only use the final responses in the scoring procedure. Another possible future research study is to investigate the proposed scoring procedure on an assessment that also contains log-file data for the MC items. In this regard, the process features can be extracted from both the MC and CR items, which could potentially further improve the performance of the proposed scoring procedure.

# Chapter 5

# Conclusion

This dissertation focused on the use of topic models within educational measurement and how these models can be used with IRMs. LSA and LDA are two commonly used topic models for assessing students' textual responses to CR items. LSA is typically used in automated scoring algorithms for CR items due to its ability to produce rubric scores that are in high-agreement with human raters. LDA is typically used in research to provide additional information about students' thinking, reasoning, and strategies when responding to CR items.

Chapter 2 presented a study that introduced a method for comparing the semantic spaces between LSA and LDA using a vector correlation known as cosine similarity. The study then used the comparison method to investigate the relationship between LSA and LDA through a simulation study and an empirical example. The simulation study used conditions that are commonly found in educational settings to generate textual responses. Once the textual responses were generated, both the LSA and LDA models were estimated and the comparison method was applied. The results found that under most conditions, the LSA and LDA models estimate similar semantic spaces. Similarly, the empirical example also showed that the semantic spaces estimated by LSA and LDA were similar.

Although the results found that LSA and LDA estimated similar semantic spaces, these two models are used for different purposes and applications. The different purposes of

these two models are partly due to how each model represents the semantic spaces. The results suggest, however, that these two models could be augmented to provide a better understanding of students' responses to CR items. For example, since LSA and LDA use the same data, LSA can be used to provide rubric scores to students and LDA can be used to provide information about the reasoning and strategies students used.

Chapter 3 presented a study that demonstrated topic models as a method for providing a substantive interpretation of a latent scale produced from an IRM. The study used an unfolding model to determine which responses to a CR item would be difficult for human raters to score accurately. The results from the unfolding model showed that there were three groups of essays: difficult-to-score below, easy-to-score, difficult-to-score above. One problem with the unfolding model, however, is that the latent scale is often hard to interpret. That is, there is it is challenging to determine the differences between the difficult-to-score below and the difficult-to-score above groups.

In this study, LDA was utilized to estimate topics for the data used to estimate the unfolding model. The topics for each unfolding group were then analyzed to identify differences among them. Results revealed that the LDA model detected distinct topic profiles for each of the three unfolding groups. These findings were then used to identify differences between the difficult-to-score below and difficult-to-score above groups. By combining topic models with unfolding models, this study offers a practical example of how to provide meaningful interpretations of the underlying latent scale.

Chapter 4 introduced a new scoring procedure for mixed-format assessments. The proposed scoring procedure uses topic models to extract features from CR items which are then used to improve the ability estimates provided by an IRM. Specifically, the new scoring procedure shows how topic models can be used to extract features from the CR items on a mixed-format assessment and how the extracted features can be integrated into the ability estimates provided by the mixed-format assessment. The scoring procedure was demonstrated

using a bootstrap method from an empirical data set. The results of the study showed that the proposed scoring procedure provided more reliable ability estimates than traditional IRM scoring methods.

In addition to providing better ability estimates, this methodology can also be used to help educators better understand the thinking and reasoning of students that are not captured in the rubric scores. Specifically, the study showed how the new scoring procedure was able to distinguish between various reasons for a student incorrectly responding to a CR item. In this regard, traditional IRM scoring methods would score the incorrect response the same way, however, the proposed scoring methodology was able to incorporate the information in the textual responses to improve the ability estimates.

The three studies presented in this dissertation showed promising results for the connection between topic models and IRMs. One intriguing finding is that topic models can be used to provide meaningful interpretation for measures obtained on assessments. Both Chapter 3 and Chapter 4 demonstrated how the results from the topic models can be used in a practical setting to better understand students' thinking and reasoning. Additionally, Chapter 2 showed the connection between LSA and LDA, which suggests that the LSA model can be used to provide scores while LDA can be used to provide interpretations of the scores.

## 5.1 Future Work and Direction

Although the three studies in this dissertation showed promising results, there are some potential limitations for using topic models in educational measurement. One such limitation is their sample dependency, that is, the estimated topics are change based on the sample used, which implies that measures obtained through topic models may lack invariance. Additionally, there are few established methods for assessing the invariance of measures obtained through topic models. Therefore, it would be beneficial to consider developing a method for measuring invariance from topic models.

The study presented in Chapter 3 demonstrated how the results from LDA can be used to provide substantive interpretation of the unfolding scale. The relationship found between the unfolding groups and the topic model results suggest that these two results could be combined to create an algorithm that flags responses that may be difficult for raters to score or an algorithm that suggests which raters should score each response. For example, the data used in Chapter 3 could be used to create an algorithm that predicts the unfolding groups (difficult-to-score below, easy-to-score, or difficult-to-score above) using the features extracted from the topic models. Then the topic model can be applied to unseen responses (similar to how automated scoring algorithms work) and the results of the topic model can be used to provide a prediction to which unfolding group a response is most likely to belong. In this regard, the results can be used to flag responses that were predicted to belong in the difficult-to-score below or the difficult-to-score above groups. In practice, this could be used to help the validity of scores provided to responses.

Additionally, unfolding models are known as preference models. This means, that raters located close to responses on the latent unfolding scale are more likely to accurately score those responses. In practice, an assessment company can calibrate raters on the unfolding scale and then apply the topic models to the unseen responses and use the algorithm to predict the response location on the unfolding scale. Then, the unseen responses can be routed to raters that have a similar unfolding location. Therefore, assessment companies can use this as a way to improve the reliability and validity of scores by human rater.

The study presented in Chapter 4 could be expanded into multiple directions. One direction would be to adapt the scoring procedure to diagnostic classification models (DCMs). In this regard, instead of estimating the ability location, the dependent variable could be the probability of having an attribute. This would be a potentially interesting direction to explore since DCMs are becoming more popular in large-scale assessments. The framework provides in Chapter 4, however, can't be directly adapted to DCMs because there are multiple

considerations. For example, DCMs have multiple attributes, therefore, there may need to be multiple scoring algorithms. Also, each CR item may measure different attributes, therefore, this would need to be explored.

Another direction would be to adapt the scoring procedure to a computerized adaptive testing (CAT) framework. This would potentially provide a framework where CAT can use CR items. It could also improve the number of items needed to accurately estimate a student's ability. There are multiple considerations, however, when adapting the scoring procedure to CAT, such as how are the CR items going to be given a rubric score, how would the item selection criteria change to include CR items, and how can topic model results be calibrated and applied to new responses.

Lastly, the scoring procedure proposed in Chapter 4 only considers features extracted from CR items. It would be interesting to investigate the performance of this scoring procedure when features are extracted from CR items and MC items. In this regard, the assessment would need to be computer-based and features would need to be extracted from CR items and process data, such as log-file data, time spent data, or eye-tracking data.

The studies presented in this dissertation provide practical uses of topic models for analyzing CR items and how these models can be used with other well-established psychometric methods, such as IRMs. With the increase of computer-based assessments, along with the accessibility of computational power, topic models have the potential to improve the quality and interpretation of measures provided by assessments.

# Bibliography


Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement*, *17*(3), 253–276.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, *4*(3).

Aubin, A.-S., St-Onge, C., & Renaud, J.-S. (2018). Detecting rater bias using a person-fit statistic: A monte carlo simulation study. *Perspectives on Medical Education*, *7*, 83–92.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. CRC Press.

Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

Blei, D., Carin, L., & Dunson, D. (2010). Probabilistic topic models. *IEEE signal processing magazine*, *27*(6), 55–65.

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

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

Boyd-Graber, J., Mimno, D., & Newman, D. (2014). Care and feeding of topic models: Problems, diagnostics, and improvements. *Handbook of mixed membership models and their applications*, *225255*.

Cardozo-Gaibisso, L., Kim, S., Buxton, C., & Cohen, A. (2019). Thinking beyond the score: Multidimensional analysis of student performance to inform the next generation of science assessments. *Journal of Research in Science Teaching*, *1*(57), 856–878.

Chang, J., & Chang, M. J. (2010). Package 'lda'.

Cho, S.-J., Brown-Schmidt, S., Boeck, P. D., & Shen, J. (2020). Modeling intensive polytomous time-series eye-tracking data: A dynamic tree-based item response model. *psychometrika*, *85*(1).

Choi, H.-J., Kwak, M., Kim, S., Xiong, J., Cohen, A. S., & Bottge, B. A. (2017). An application of a topic model to two educational assessments. *The Annual Meeting of the Psychometric Society*, 449–459.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, *17*(1), 61–84.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5*(1).

Duong, E., Mellom, P., & Hixon, R. (2019). Using topic modeling to analyze the effects of instructional conversation on 3rd grade students' writing.

Eckes, T. (2005). Examining rater effects in testdaf writing and speaking performance assessments: A many-facet rasch analysis. *Language Assessment Quarterly: An International Journal*, *2*(3), 197–221.

Engelhard, G. (2012). Examining rating quality in writing assessment: Rater agreement, error, and accuracy. *Journal of applied measurement*, *13*, 321–35.

Engelhard Jr, G. (1997). Constructing rater and task banks for performance assessments. *Journal of outcome measurement*, *1*(1), 19–33.

Engelhard Jr, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted rasch model. *Journal of educational measurement*, *31*(2), 93–112.

Engelhard Jr, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, *33*(1), 56–70.

Engelhard Jr, G. (2013). *Invariant measurement: Using rasch models in the social, behavioral, and health sciences*. Routledge.

Ercikan, K., & Pellegrino, J. W. (2017). *Validation of score meaning for the next generation of assessments: The use of response processes*. Taylor & Francis.

Evangelopoulos, N., Zhang, X., & Prybutok, V. R. (2012). Latent semantic analysis: Five methodological recommendations. *European Journal of Information Systems*, *21*(1), 70–86.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes*, *25*(2-3), 285–307.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, *1*(2), 939–944.

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, *22*(4), 521–563.

Golub, G. H., & Reinsch, C. (1971). Singular value decomposition and least squares solutions. *Linear algebra*, *2*, 134–151.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, *36*(2), 193–202.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl_1), 5228–5235.

Grün, B., & Hornik, K. (2011). Topicmodels: An r package for fitting topic models. *Journal of statistical software*, *40*, 1–30.

Han, K.-X., Chien, W., Chiu, C.-C., & Cheng, Y.-T. (2020). Application of support vector machine (svm) in the sentiment analysis of twitter dataset. *Applied Sciences*, *10*(3), 1125.

He, Q., & Davier, M. v. (2015). Identifying feature sequences from process data in problem-solving items with n-grams. In *Quantitative psychology research* (pp. 173–190). Springer.

He, Q., Veldkamp, B. P., Glas, C. A., & Van Den Berg, S. M. (2019). Combining text mining of long constructed responses and item-based measures: A hybrid test design to screen for posttraumatic stress disorder (ptsd). *Frontiers in psychology*, *10*, 2358.

Kakkonen, T., Myller, N., & Sutinen, E. (2006). Applying latent dirichlet allocation to automatic essay grading. *Advances in Natural Language Processing: 5th International Conference on NLP, FinTAL 2006 Turku, Finland, August 23-25, 2006 Proceedings*, 110–120.

Kim, H., Howland, P., Park, H., & Christianini, N. (2005). Dimension reduction in text classification with support vector machines. *Journal of machine learning research*, *6*(1).

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*.

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of irt linking methods for mixed-format tests. *Applied Measurement in Education*, *19*(4), 357–381.

Kim, S., Walker, M. E., & McHale, F. (2010). Comparisons among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement*, *47*(1), 36–53.

Kwak, M. (2019). *Parameter recovery in latent dirichlet allocation (lda): Potential utility of lda in formative constructed response assessment* (Doctoral dissertation). University of Georgia.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259–284.

LaVoie, N., Parker, J., Legree, P. J., Ardison, S., & Kilcullen, R. N. (2020). Using latent semantic analysis to score short answer constructed responses: Automated scoring of the consequences test. *Educational and Psychological Measurement*, *80*(2), 399–414.

Li, B., & Han, L. (2013). Distance weighted cosine similarity measure for text classification. *Intelligent Data Engineering and Automated Learning–IDEAL 2013: 14th International Conference, IDEAL 2013, Hefei, China, October 20-23, 2013. Proceedings 14*, 611–618.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. IAP.

Mardones-Segovia, C., Choi, H.-J., Hong, M., Wheeler, J. M., & Cohen, A. S. (2022). Comparison of estimation algorithms for latent dirichlet allocation. *Quantitative Psychology: The 86th Annual Meeting of the Psychometric Society, Virtual, 2021*, 27–37.

Mardones-Segovia, C., Wheeler, J. M., Choi, H.-J., & Cohen, A. S. (2021). Model selection for latent dirichlet allocation model selection for latent dirichlet allocation with small number of topics.

Mcauliffe, J., & Blei, D. (2007). Supervised topic models. *Advances in neural information processing systems*, *20*.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational researcher*, *23*(2), 13–23.

Molenaar, D., Oberski, D., Vermunt, J., & De Boeck, P. (2016). Hidden markov item response theory models for responses and response times. *Multivariate behavioral research*, *51*(5), 606–626.

Mozgovoy, M., Kakkonen, T., & Cosma, G. (2010). Automatic student plagiarism detection: Future perspectives. *Journal of Educational Computing Research*, *43*(4), 511–531.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet rasch measurement: Part i. *Journal of applied measurement*, *4*(4), 386–422.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*(4), 341–384.

Ponweiser, M. (2012). Latent dirichlet allocation in r.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 111–163.

Ramesh, D., & Sanampudi, S. K. (2021). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, 1–33.

Read, B., Francis, B., & Robson, J. (2005). Gender,'bias', assessment and feedback: Analyzing the written assessment of undergraduate history essays. *Assessment & Evaluation in Higher Education*, *30*(3), 241–260.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing writing*, *15*(1), 18–39.

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., & Steyvers, M. (2010). Learning author-topic models from text corpora. *ACM Transactions on Information Systems (TOIS)*, *28*(1), 1–38.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological bulletin*, *88*(2), 413.

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, *81*(10.5555), 26853.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.

Samejima, F. (1997). Graded response model. In *Handbook of modern item response theory* (pp. 85–100). Springer.

Schofield, A., Magnusson, M., Thompson, L., & Mimno, D. (2017). Understanding text pre-processing for latent dirichlet allocation. *Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics*, *2*, 432–436.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*, 333–343.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a united states demonstration. *Assessing Writing*, *20*, 53–76.

Shermis, M. D., & Burstein, J. (2013). Handbook of automated essay evaluation. *NY: Routledge*.

Shin, J., Guo, Q., & Gierl, M. J. (2019). Multiple-choice item distractor development using topic modeling approaches. *Frontiers in psychology*, *10*, 825.

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.

Steinberger, J., Jezek, K., et al. (2004). Using latent semantic analysis in text summarization and summary evaluation. *Proc. ISIM*, *4*(93-100), 8.

Tang, X., Wang, Z., He, Q., Liu, J., & Ying, Z. (2020). Latent feature extraction for process data via multidimensional scaling. *psychometrika*, *85*(2), 378–397.

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(3), 611–622.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*, 29–44.

Uto, M. (2019). Rater-effect irt model integrating supervised lda for accurate measurement of essay writing ability. *International Conference on Artificial Intelligence in Education*, 494–506.

Wall, M. E., Rechtsteiner, A., & Rocha, L. M. (2003). Singular value decomposition and principal component analysis. *A practical approach to microarray data analysis*, 91–109.

Wang, J., & Engelhard Jr, G. (2019a). Conceptualizing rater judgments and rating processes for rater-mediated assessments. *Journal of Educational Measurement*, *56*(3), 582–609.

Wang, J., & Engelhard Jr, G. (2019b). Exploring the impersonal judgments and personal preferences of raters in rater-mediated assessments with unfolding models. *Educational and psychological measurement*, *79*(4), 773–795.

Wang, J., Engelhard Jr, G., Raczynski, K., Song, T., & Wolfe, E. W. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing*, *33*, 36–47.

Wang, J., Engelhard Jr, G., & Wolfe, E. W. (2016). Evaluating rater accuracy in rater-mediated assessments using an unfolding model. *Educational and Psychological Measurement*, *76*(6), 1005–1025.

Wang, L., & Wan, Y. (2011). Sentiment classification of documents based on latent semantic analysis. *International Conference on Computer Education, Simulation and Modeling*, 356–361.

Wang, W.-C., Su, C.-M., & Qiu, X.-L. (2014). Item response models for local dependence among multiple ratings. *Journal of Educational Measurement*, *51*(3), 260–280.

Watanabe, S., & Opper, M. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, *11*(12).

Weigle, S. C., Yang, W., & Montee, M. (2013). Exploring reading processes in an academic reading test using short-answer questions. *Language Assessment Quarterly*, *10*(1), 28–48.

Wheeler, J. M., Cohen, A., & Wang, S. (2021). Comparison of latent semantic analysis and latent dirichlet allocation. *Paper presented at the Annual Meeting of the National Council on Measurement in Education (Virtual)*.

Wheeler, J. M., Cohen, A. S., & Wang, S. (under revision). A comparison of latent semantic analysis and latent dirichlet allocation in educational measurement. *Journal of Educational and Behavioral Statistics*.

Wheeler, J. M., Cohen, A. S., Xiong, J., Lee, J., & Choi, H.-J. (2021). Sample size for latent dirichlet allocation of constructed-response items. In *Quantitative psychology* (pp. 263–273). Springer.

Wheeler, J. M., Engelhard, G., & Wang, J. (2022a). Exploring rater accuracy using unfolding models combined with topic models: Incorporating supervised latent dirichlet allocation. *Measurement: Interdisciplinary Research and Perspectives*.

Wheeler, J. M., Engelhard, G., & Wang, J. (2022b). Exploring rater accuracy using unfolding models combined with topic models: Incorporating supervised latent dirichlet allocation. *Measurement: Interdisciplinary Research and Perspectives*, *20*(01), 34–46. https://doi.org/10.1080/15366367.2021.1915094

Wheeler, J. M., Raczynski, K. R., Cohen, A. S., & Engelhard Jr, G. (2022). Using topic models to understand rater-mediated writing assessments. *The Journal of Experimental Education*, 1–20.

Wheeler, J. M., Wang, S., Tan, Y., & Cohen, A. (2022). Textual data as process data: A new scoring procedure for mixed-format assessments.

Wheeler, J. M., Wang, S., Tan, Y., & Cohen, A. S. (under revision). Textual data as process data: A new scoring procedure to improve ability estimation for mixed-format assessments. *Psychometrika*.

Wheeler, J. M., Xiong, J., Mardones-Segovia, C., Choi, H.-J., & Cohen, A. S. (2022). An investigation of prior specification on parameter recovery for latent dirichlet allocation of constructed-response items. *Quantitative Psychology: The 86th Annual Meeting of the Psychometric Society, Virtual, 2021*, 203–215.

Wilbur, W. J., & Sirotkin, K. (1992). The automatic identification of stop words. *Journal of information science*, *18*(1), 45–55.

Wild, F. (2007). An lsa package for r. *Proceedings of the 1st International Conference on Latent Semantic Analysis in Technology Enhanced Learning (LSA-TEL'07)*, 11–12.

Wild, F. (2016). *Learning analytics in r*. Springer.

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, *26*(3), 283–306.

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, *31*(3), 31–37.

Xiong, J., Choi, H.-J., Kim, S., Kwak, M., & Cohen, A. S. (2020). Topic modeling of constructed-response answers on social study assessments. *Quantitative Psychology: 84th Annual Meeting of the Psychometric Society, Santiago, Chile, 2019 84*, 263–274.

Xiong, J., Wheeler, J. M., Choi, H.-J., Lee, J., & Cohen, A. S. (2021). An empirical study of developing automated scoring engine using supervised latent dirichlet allocation. In *Quantitative psychology* (pp. 429–438). Springer.

Zhang, S., Wang, Z., Qi, J., Liu, J., & Ying, Z. (2021). Accurate assessment via process data. *arXiv preprint arXiv:2103.15034*.