Characterizing the Organizational Identities of California Community Colleges: A Comparison of Manual and Machine Learning Methods.

Fowler, Caleb L

Characterizing the Organizational Identities of California Community Colleges: A Comparison of Manual and Machine Learning Methods.

By

CALEB L. FOWLER, JR.
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF EDUCATION

in

Educational Leadership

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Cassandra Marie Doll Hart, Chair

Michal Kurlaender

Megan Welsh

Committee in Charge

2023

ABSTRACT

It is becoming ever more challenging to address the changing landscape faced by California Community Colleges. Change agents strive to implement needed interventions but are often disappointed in the outcomes. However, the literature shows that the odds of successful change increase when interventions align with the organization's culture. The problem is how do change agents identify elements of a college's culture?

This study explores that question by using different methodological approaches to extract elements of college identity from documents that plausibly express key college priorities and values. Using the organizational identity framework, elements of college identity are conceptualized as latent themes embedded within artifacts produced by individual colleges - in this case, Institutional Self Evaluation Reports (ISERs) used for accreditation.

To generate the broadest possible picture, my dissertation uses two different methods: a machine-learning content analysis method and a manual semantic analysis method. Using Latent Dirichlet Allocation, the computer extracted 5 latent topics from 25 Community College ISERs corpus. Concurrently using Dedoose, manual semantic analysis was conducted with vivo coding of the same corpus. Latent topics emerged by analyzing both sets of semantic topics separately.

Several results emerged from the study: it confirmed that ISERs are a suitable data set for topic extraction, and both methods can extract latent themes. However, the themes did not overlap as much as anticipated. The automated method extracted topics with an institutional lens focused on specific processes, while the manual method used a broader perspective, focusing on outcome values and concepts.

The results suggest that both methods, used together, can provide a more comprehensive picture than either method used alone. The study concludes with several recommendations that practitioners may find useful.

Keywords: Community College, Institutional Self Evaluation Report, Machine Learning, Organizational Identity, Semantic Analysis, Content Analysis, Topic Extraction

TABLE OF CONTENTS

DEDICATION

To all those brave souls who dare to question the prevailing paradigms, the dreamers who paint a picture of a more promising world, and the tenacious individuals who drive our beloved institutions forward, this dissertation is devoted to you. To my wife, Dorine, and my daughters, Elora, Madelena, and Angelica, I owe an immeasurable debt of gratitude for their unwavering love and encouragement throughout these past several years. This work stands as a testament to the collective support of these remarkable champions, who have been my guiding light on this academic journey.

ACKNOWLEDGMENTS

I extend my sincere gratitude to Folsom Lake College for providing me with the space and environment to pursue and complete this dissertation. This assistance was invaluable in making this academic journey possible. Additionally, I am thankful to the UC Davis CANDEL program for its unwavering focus on student success. Their dedication to excellence has been a motivating force throughout my studies. This work would not have been possible without the advice and guidance of my dissertation chair, Dr. Cassandra Marie Doll Hart. Thank you for your patience and invaluable assistance. I would also like to thank the members of my dissertation committee, Dr. Michal Kurlaender and Dr. Megan Welsh, for their hard work and support.

Furthermore, I am deeply grateful for the honest feedback of my colleagues Paula Levin, Dan Considine, and Bernadette Anayah, who have been a constant source of ideas and a willing ear when I needed someone to listen. Their valuable suggestions and insights have undoubtedly enriched this work and my overall academic experience. I want to thank my sister, Liana, and her husband, Sean, and my nephews, Bowie and Hendrix. I know I can always count on them and am grateful for their support. Above all, I owe a debt of gratitude to my mother, Lynne Fowler, and my father, Caleb Fowler. Their belief in my abilities has been the bedrock of my academic expedition. They have not only encouraged me to pursue my dreams but also shown me the true meaning of grace under pressure. Their love and guidance have shaped me into the person I am today, and I am eternally thankful for that.

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION AND PURPOSE

"We can see the identity of an organization as the sum of the policy and procedures it enacts."
- Marlene Fiol (Fiol, 1991)

"Community colleges are having an identity crisis" (AACRAO, 2018, p. 1). The community college landscape has become substantially more complex over the last decade. Colleges have implemented changes to adapt to the evolving landscape, but the results could have been better. Yet, change agents can be "more successful when they adopt strategies that fit the culture of the institution for which they are trying to make [a] change" (Kezar, 2018, p. xi). However, change agents must understand the institution's values before deciding on an intervention (Kezar, 2018, p. x). Organizational Identity (OI) is an area of study that seeks to understand how organizations answer the question of "who are we as an organization?" (Albert & Whetten, 1985); OI aims to identify and articulate the organization's values to facilitate understanding. All subsequent activities flow from the college's understanding of who they are.

A number of studies have characterized elements of college identity by analyzing mission and vision statements (Ayers, 2002; Chen et al., 2020; Davies & Glaister, 1996; Kuenssberg, 2011; Perez-Encinas et al., 2021; Seeber et al., 2019; Sun et al., 2019; Wang et al., 2007). While most of these papers use manual qualitative coding of mission and vision statements to uncover elements of organizational identity (Ayers, 2002; Davies & Glaister, 1996; Kuenssberg, 2011; Perez-Encinas et al., 2021; Seeber et al., 2019; Wang et al., 2007), a handful of studies have begun to explore machine learning methods to aid in coding these statements (Chen et al., 2020; Sun et al., 2019). While the latter method holds promise for scaling up this area of research by examining mission and vision statements in a large dataset—such as in the IPEDS or a corpus of college strategic plans — it is unknown whether the identity themes extracted with automated and manual methods are congruent. Each method using the same data may generate a unique set of themes for the same college.

Moreover, alternate documents may better help decode themes of organizational identities. Mission and vision statements are not very detailed, nor are they very long. A larger, deeper document containing similar content should provide more sophisticated results. Institutional Self Evaluation Reports (ISERs) — self-evaluation reports that colleges write for accreditation purposes — are larger, richer documents that generally include college mission and vision statements, as well as initiatives that colleges believe display how they interact with their missions and visions. These reports may provide additional data to help uncover differences in organizational identity across colleges.

This mixed-methods dissertation study aimed to fill these gaps in the literature. Specifically, my study compared two topic extraction techniques — a manual content analysis qualitative method and an automated machine-learning semantic analysis method. I used the data from Institutional Self Evaluation Reports (ISER) produced by the colleges for reaccreditation purposes. Colleges must address the topics the accrediting agency specifies as accreditation standards and document compliance with these standards in the ISER (Accrediting Commission for Community and Junior Colleges, 2022). Consequently, all ISERs contain a standard set of information in a standard sequence, addressing a standard set of questions. Standard IA of the ISER, for example, addresses the college mission and should contain the information needed for identifying identity themes.

Research Questions

This inquiry aimed to provide additional information for college leaders to improve decision-making. My study addressed the following research questions:

1. What were the main identity themes of CA community colleges as revealed by automated topic extraction analysis of ISER Standard IA?
2. What were the main identity themes of CA community colleges as determined by manual topic extraction analysis of ISER Standard IA?
3. What were the similarities between the topics extracted by automated and manual topic extraction methods?
4. What were the differences between the topics extracted by automated and manual topic extraction methods?

I used a mixed-method design and applied different methods to the same corpus of documents to attempt to extract latent concepts of organizational identity. This allowed me to compare the resulting data sets and understand the impact of research methods on generating organizational identity themes. This resulted in a deeper understanding of the problem (Creswell & Creswell, 2020). My overall method was thematic analysis, an overarching framework that binds together the methodologies used to compare the dataset (Braun & Clarke, 2022). The process I used to accomplish this comparison of methods was as follows: 1) I applied automated semantic analysis — a machine learning method commonly used in other social science disciplines — to Standard IA of the ISERs to extract key topics; 2)I manually coded Standard IA of the same ISERs to extract key semantic topics using manual Content Analysis (Merriam & Tisdell, 2016, p. 179); 3) I derived two groups of latent identity themes by analyzing the groups of semantic topics produced in steps one and two; and 4) I described the differences and similarities of elements of organizational identity extracted from these two methods. The corpus was derived from a random sample of 25 community college ISERs downloaded from the appropriate college website. Table 1 summarizes this process.

Table 1

Process Summary

	Research Question 1	**Research Question 2**	**Research Question 3**	**Research Question 4**
Question	What were the main identity themes of CA community colleges as revealed by automated topic extraction analysis of ISER Standard IA?	What were the main identity themes of CA community colleges as determined by manual topic extraction analysis of ISER Standard IA?	What were the similarities between the topics extracted by automated and manual topic extraction methods?	What were the differences between the topics extracted by automated and manual topic extraction methods?
Purpose and Design Rational	Create a baseline of identity themes for comparison using automated quantitative topic extraction methods.	Create a baseline of identity themes for comparison using manual qualitative topic extraction methods.	Understand how the two semantic methods yield similar results.	Understand how the two semantic methods yield dissimilar results.
Methodology	Semantic coding is performed with a machine learning algorithm, followed by latent coding by a human researcher.	Qualitative semantic, and latent coding are produced manually by a human researcher.	Manual comparison of results of first to research questions, research, notes, and other analysis documents.	

The remainder of my study is organized as follows: first, I present the relevant literature in which my study is grounded; second, I describe my research design; third, I discuss the results of my study; and finally, I discuss the implications of my findings.

CHAPTER TWO: REVIEW OF THEORY AND PRIOR LITERATURE

"The man who does not read has no advantage over the man who cannot read."
- Mark Twain

I contextualize my research by examining academic conversations related to my topic. First, I discuss the theoretical framework which guides this study — Organizational Identity (OI). I then review the literature on using mission and vision statements in the academy as evidence of organizational identity and the literature on college accreditation reports as a dataset. These scholarly works are the foundation upon which I build my research questions. Part of my dissertation also relies on comparing qualitative manual and quantitative machine-based methods to extract topics. Therefore, I also review the literature on machine learning and manual educational content analysis approaches. I conclude with a discussion of the gaps in the literature and their relation to this project.

Organizational Identity Defined

Albert and Whetten's (1985; 2006) Organizational Identity (OI) framework is the conceptual backbone of this study. Fundamentally, OI consists of three central ideas: (a) identity is a property of the organization articulated by its members; (b) identity is composed of organizational elements believed to be central, enduring, and distinctive (Albert & Whetten, 1985; Gioia & Hamilton, 2016, p. 25; Whetten, 2006); and (c) identity-based membership claims occupy specific social and geographic spaces. Whetten (2006) felt these ideas were more prominent when the organization was dealing with "profound fork in the road choices [which have] the potential to alter the collective understanding of 'who we are as an organization'" (p.221).

Organizational Identity is based on two assumptions: 1) organizations are more than social collectives, and 2) identity is an unobservable subjective state and may be inferred only from effects or consequences. This situates every organization within a self-determined and self-defined social space. Proper classification is critical to recognition, which is crucial to identity — consequently, possessing a strong OI confers several organizational advantages. Organizations with strong identities may be more adaptable in times of instability and environmental change (Gioia et al., 2000; Ran & Golden, 2011) because identity serves as both a foundation to anchor decisions (e.g., for community colleges: 'our goal is to serve students, not make a profit') and a scaffold to guide decision making ('how does this serve our students?') (Kezar, 2018). OI has also proven useful for studying organizational dynamics: conceptualizing how and why individuals respond to organizational threats (Petriglieri & Devine, 2016, p. 239), understanding the underlying rationale for specific strategic decisions (Nag et al., 2007), identifying specific attributes of organizational structure (Clark et al., 2010; Corley & Gioia, 2004), developing insight on the relationship between employees and their organizations (Dutton et al., 1994; Fiol, 2002; Glynn, 2000; Pratt, 2000), and even determining how predisposed toward adopting innovations an organization may be (Anthony & Tripsas, 2016, p. 417). A fundamental assumption across these studies is that "… identity figures prominently in what actions organizations take" (Levin, 2001, p. 10). Organizational Identity is an emergent property resulting from the outcomes of past decisions and responses to organizational challenges. I posit that this leaves behind observable artifacts in the form of policies, procedures, and processes.

Industry, organizational form, and accrediting bodies lend social identity and legitimacy to the organization (Kezar, 2018). Colleges tend to look toward their peers and respond in kind to novel situations (DiMaggio & Powell, 1983). Thus, they do not want to appear too different for fear of losing their identity altogether ('Oh?, they are a college?') (Basko, 2022). Accreditation, in particular, can drive organizational homogeneity because colleges strive to comply with identical accreditation standards. However, research suggests colleges are more effective when they implement unique strategies based on the colleges' strengths (DiMaggio & Powell, 1983; Kezar, 2018) and can legitimately differentiate themselves from each other (Basko, 2022).

Figure 1

Situating the OI Framework

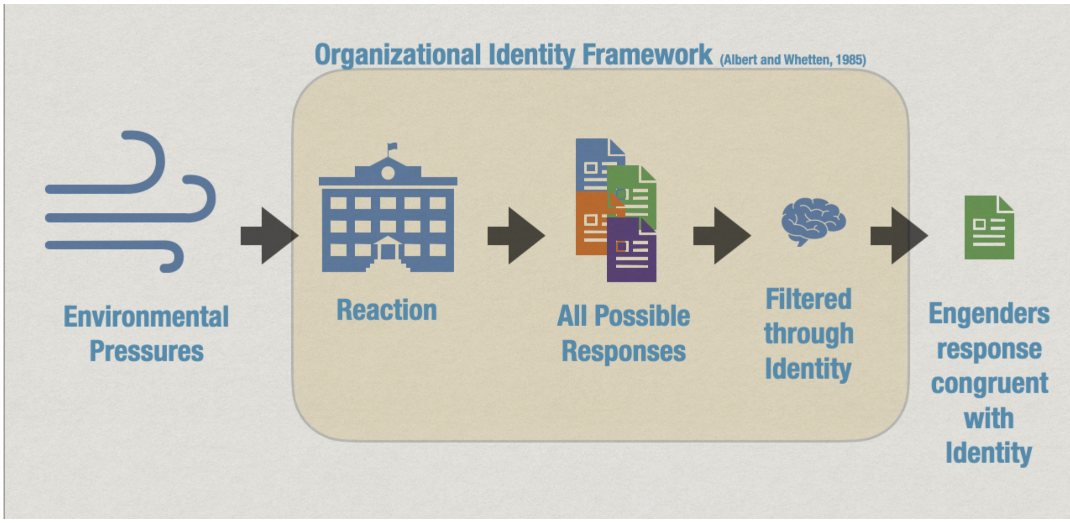

Organizational Identity Framework (Albert and Whetten, 1985)
Environmental Pressures
Reaction
All Possible Responses
Filtered through Identity
Engenders response congruent with Identity

Figure 1 above elucidates why the organizational identity framework (Albert & Whetten, 1985; Whetten, 2006) is helpful in this study because it illustrates how researchers can identify a slippery concept like identity through the cognitive artifacts an organization produces. Assume that environmental pressures act on an organization. This action leads to some response or reaction. Often, there are several possible responses that the organization can take. OI suggests that all possible actions are filtered through the idea of who the organization thinks it is, that is, the organization's identity. It also assumes that organizations respond in a characteristic way that is true to their identity. Cognitive artifacts often record these espoused responses (committee minutes, plans, mission and vision statements, ISERs, etc.). This study explores actions memorialized in a specific cognitive artifact — college ISERs — whereby colleges enumerate actions they have taken that demonstrate commitments flowing from their organizational identity.

Mission and Vision Analysis in Higher Education

We can expect elements important to an organization's identity to manifest in efforts that are important to the organization. College mission and vision statements are important to the college and, therefore, are an excellent choice to use as data for extracting latent OI themes. Collectively known as institutional vision statements, these documents identify aspirations, establish commitments, and reinforce expectations (Fox et al., 2003; Pekarsky, 2007). First becoming popular in Corporate America (Drucker, 1974; Peters & Waterman, 1982), institutional vision statements later found acceptance in the academy (Birnbaum & Snowdon, 2003; Davies & Glaister, 1996).

Despite widespread acceptance, mission and vision statements have not escaped controversy. For example, researchers critical of mission and vision statements suggest they say the same things and lack real meaning. Newsom and Hayes (1991) summarized this work when they wrote, "[mission statements are] full of honorable verbiage signifying nothing," possessing an overall sameness (Newsom & Hayes, 1991), rarely containing aspirational statements (Morphew & Hartley, 2006) with slight variance in terms (Firmin & Gilson, 2009), and containing "vague and vapid goals" (Chait, 1979, p. 36).

Other researchers contend that the homogeneity of mission statements is sensible if the primary purpose is to legitimize rather than differentiate institutions. Ayers (2015) examined how discourses reflect institutional logics — sensemaking frames that provide an understanding of what is legitimate, reasonable, and effective within a given organizational context — and illuminated the relationship between the community college and its environment. For instance, he observed an increased frequency in the use of the word 'degree' in community college statements from 2004 to 2012 to a "statistically exceptional" (Ayers, 2015, p. 200) amount, which the author suggested reflected the colleges flagging their degree completion goals as a legitimization strategy. Morphew and Hartley (2006) theorize that mission and vision statements are legitimizing documents because they demonstrate that the college understands the 'rules of the game' — and one of the rules is that you have to have a mission statement if you are a college. Gioia and Hamilton (2016) spoke directly to the institutional perspective of the OI framework by examining the categorical self-descriptors colleges use. These are the terms a college would use to describe itself — often in relation to other social actors (Whetten & Mackey, 2002, p. 396). They suggest that a significant part of a college's identity is derived from its membership claims in a specific social category ('we are a liberal arts school,' for example). In these perspectives, it is unsurprising when the mission statements of colleges reflect similarities, given that colleges use them to demonstrate their organizational legitimacy.

Other work contradicts the idea that mission statements reflect an overarching similarity, exploring variation in the use of college institutional vision statements and their themes. For instance, Ruef and Nag (2015) developed a new college classification scheme based on emergent categorical descriptors from the text in the colleges' vision statements. Other researchers relate specific elements of mission statements to either outcomes or college characteristics. Palmer and Short (2008), investigating the mission statements of business schools, found considerable variance in word choice and additionally observed that the more complete a mission statement was, the better performing the college. Another set of studies has explored how the language of the mission and vision statement varied according to some independent variable such as the size of the institution (Cortés Sánchez, 2018; Cortés-Sánchez, 2017; Palmer & Short, 2008), university reputation (Seeber et al., 2019), sector (public or private) (Efe & Ozer, 2015), or the impact of geographic location (Bayrak, 2020). All of these studies use variations in mission and vision statements to extract themes and make comparisons between institutions.

Several researchers have extracted latent themes from college or university mission and vision statements using methods similar to those proposed in this study.[1] Ayers (2002) conducted a content analysis of the mission statements of 102 community colleges in the southern United States and identified seven themes: (a) access, (b) workforce and economic development, (c) comprehensive programs, (d) quality excellence, (e) responsiveness to needs,

[1] While I highlight the most relevant example here, I refer the reader to other studies by Hladchenko (2013) and Kuenssberg (2011) that may also be of interest.

(f) specified service area, and (g) diversity. Furthermore, Ayers (2002) identified that the access theme (a) occurred significantly more than either the service area (f) or diversity (g) themes. There were no significant differences in the proportion of the access theme to the other four remaining themes.

Seeber et al. (2019) conceptualized mission statements from 123 universities in the United Kingdom as identity narratives. Applying concepts from the organizational identity literature to these narratives allowed them to hypothesize the factors that affect their institutional vision statements. They determined that claims in mission and vision statements fell into one of four categories: (a) declarations about the university's intended goals, (b) declarations about the university's intended means, (c) claims regarding the university's intended customers either within a specific geographical area (Bay Area, for example) or a given demographic (students, parents, etc.), or a (d) quantifier regarding the university's activities or services (such as efforts to differentiate themselves by claims of 'accessibility,' 'excellence,' or 'competence').

Wang et al. (2007) examined the mission statements of 34 four-year colleges and 68 two-year colleges in Texas to determine how similar or dissimilar the themes extracted from each group were. Fifteen reoccurring themes were identified: leadership, citizenship, cultural diversity, lifelong learning, excellence in teaching and research, creativity, critical thinking, academic achievement, collaboration and partnership, vocational and technical skills, access to higher education, academic readiness and skill development, student services, community focus, and technology. Quantitative analysis revealed statistically significant differences between the two Texas higher education institutions groups on eight of the 15 themes (Wang et al., 2007). A comparison of the percentage of term occurrence suggested that four-year institutions focused on the six themes: leadership, citizenship, cultural diversity, excellence in teaching and research, creativity, and academic achievement, while two-year institutions focused on the two terms: 'vocational and technical skills' and 'development of academic readiness skills.' In summary, mission and vision statements have been valuable research data sources to extract identity themes.

However, despite their utility, there are several issues with mission and vision statements as data sources. Primarily, they suffer from the problem of being relatively brief. Generally, they are small, between one and three sentences, and rarely exceed 100 words (Tsang, 2020), with an average length of 29 words (Brandall, 2018). This can yield very little data, so incorporating additional text related to the mission and vision statements may be helpful. This is especially true in machine learning, as the algorithms perform much better with more extensive texts (Lewis, 2017). This study addresses this problem by looking at the additional text the colleges are prompted to write to support how their mission and vision statements are enacted in college practices, specifically, responses to Standard 1A of the Institutional Self Evaluation Reports.

In addition, community colleges are not required to produce vision statements, meaning that not all colleges generate institutional vision statements that can contribute to the analysis. However, accrediting bodies generally require self-reports, and the accrediting agency that governs the colleges used in this study requires a discussion of essential elements of the college's mission (appearing in Standard 1A of the Institutional Self Evaluation Report). Furthermore, the ACCJC suggests colleges write six pages addressing Standard IA (Accrediting Commission for Community and Junior Colleges, 2022, p. 14). ACCJC accreditation reports provide data similar to what is found in a mission and vision statement, often including design rationale as well. This makes Standard IA a richer data set for determining elements of organizational identity than mission and vision statements that have been the focus of much of the past research.

Several conclusions may be drawn from this literature. Researchers have used mission and vision statements to extract latent themes from various colleges. This research has also explored other hard-to-examine topics, such as reputation, similarity with other colleges, and abstract concepts, such as 'education for all' using mission and vision statements as data. However, extracting diverse identity themes may prove difficult because mission and vision statements are relatively small. Using the larger text of Standard IA would open the study to a richer data set and provide more diverse latent themes.

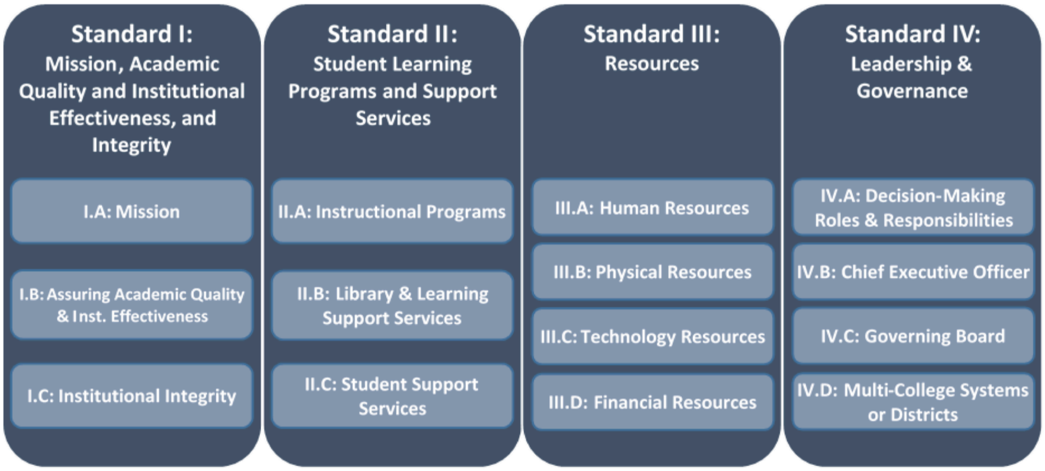
Accreditation Reports

To improve the quality of the extracted topics, this study incorporates additional text using accreditation reports, which contain mission statements and often vision statements, and surrounding text elaborating on the themes and design rationale in the mission statement. In the California Community College system, accreditation reports, also known as Institutional Self Evaluation Reports (ISERs), are produced as part of the seven-year accreditation process. The college typically starts writing the ISER a year or so before the college's periodic accreditation cycle starts. The accrediting agency creates standards that focus on and guide the creation of the ISER (Accrediting Commission for Community and Junior Colleges, 2022). Once the document is written, it is submitted to the accreditor for review. Representatives of the accrediting agency then verify the content of the report. The report they submit is used along with the ISER in determining the accreditors' final decision on accreditation status (Accrediting Commission for Community and Junior Colleges, 2022). Accreditation status is vital for the college as colleges that lose their accredited status often experience numerous struggles for survival (Burnett, 2020).

The accrediting agency promulgates the accreditation standards, but how the college addresses them represents a shared understanding between the college staff and its place in the world. That shared understanding creates a frame of reference (Argyris & Schön, 1996) that influences the choice of what projects and programs will be implemented. College staff scan, filter, and select (Weick, 2001) elements from the environment compatible with "who they believe they are as an organization" (Albert & Whetten, 1985; Whetten, 2006). This self-referential framework guides and focuses staff perceptions and behaviors. The standards and accreditation process serve as a stimulus, but the colleges' organizational identities are assumed to influence how the college manifests the concepts in the standard. The ISER is documentation attesting to compliance with accreditation standards. It aggregates data supporting adherence to standards enumerated by the ACCJC and provides a place for the college to discuss why it made its choices.

Figure 2

ACCJC Accreditation Standards (2014)



Standard I: Mission, Academic Quality and Institutional Effectiveness, and Integrity	Standard II: Student Learning Programs and Support Services	Standard III: Resources	Standard IV: Leadership & Governance
I.A: Mission	II.A: Instructional Programs	III.A: Human Resources	IV.A: Decision-Making Roles & Responsibilities
I.B: Assuring Academic Quality & Inst. Effectiveness	II.B: Library & Learning Support Services	III.B: Physical Resources	IV.B: Chief Executive Officer
I.C: Institutional Integrity	II.C: Student Support Services	III.C: Technology Resources	IV.C: Governing Board
		III.D: Financial Resources	IV.D: Multi-College Systems or Districts

Note: (Accrediting Commission for Community and Junior Colleges, 2022)

Figure 2 is an abbreviated listing of ACCJC accreditation standards. This study will examine only Standard IA: Mission because that standard directly deals with the mission and vision of the university. Below are the four sub-standards comprising this standard that articulate the aspects of the mission and vision that will receive attention from the college:

> I.A.1. The mission describes the institution’s broad educational purposes, its intended student population, the types of degrees and other credentials it offers, and its commitment to student learning and student achievement (Accrediting Commission for Community and Junior Colleges, 2022, p. 35).
>
> I.A.2. The institution uses data to determine how effectively it is accomplishing its mission, and whether the mission directs institutional priorities in meeting the educational needs of students (Accrediting Commission for Community and Junior

> Colleges, 2022, p. 36).
>
> I.A.3. The institution's programs and services are aligned with its mission. The mission guides institutional decision-making, planning, and resource allocation and informs institutional goals for student learning and achievement.
>
> I.A.4. The institution articulates its mission in a widely published statement approved by the governing board. The mission statement is periodically reviewed and updated as necessary (Accrediting Commission for Community and Junior Colleges, 2022, p. 37).

Colleges meet this standard by including their mission and vision statements and providing additional commentary, design rationale, and summaries of supporting documentation. This is much larger than typical institutional vision statements and should give a more complete picture of the semantic and latent themes around identity statements important to the college.

Accreditation, in general, has received little attention in the literature, but some studies have examined the intersection between college performance and accreditation outcomes. Sodhi (2016) reviewed the impact of staff perception of the accreditation process and compliance with Standard I.B — Assuring Academic Quality and Institutional Effectiveness. She determined that negative perceptions regarding the ACCJC hinder efforts to comply with Standard IB. This is also noteworthy because it is a rare paper focusing on a specific accreditation standard and examining the same accreditor used in this study. Other researchers focused on the impact of the outcome of the accreditation process. One consistent finding is the weak association between accreditation status, student outcome variables, and student performance (Djeukeng, 2014; Hossain et al., 2019; Rybinski, 2020; Serafin, 2014; Theule, 2012). This study extends this modest literature by using accreditation reports as a data set to investigate organizational identity.

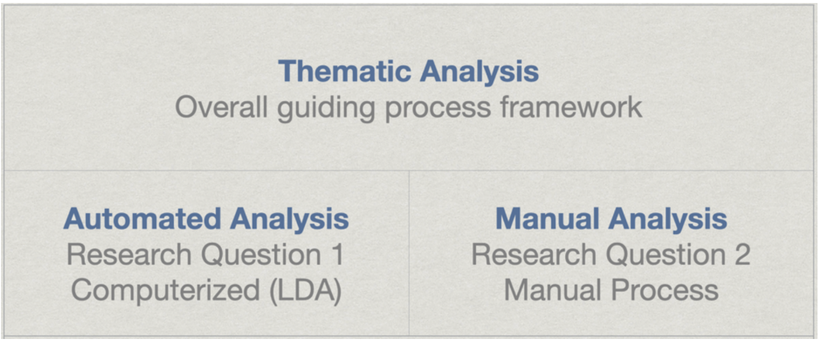
Textual Analysis in Education

One of the goals of this project is to compare the performance of manual coding and automated machine-learning models in identifying themes relevant to the organizational identity of community colleges. Machine learning methods are much less established tools for analyzing the organizational identity of colleges than manual, qualitative methods. To this end, I provide an overview of three methodologies relevant to this study and describe how these methods are used together. While it is important to provide a general overview of these methods in this section to situate the study's contribution, I provide specific details on the procedures used to carry out each method in the Methods section.

Figure 3 below shows the three interrelated methods. The overall methodology for this project is thematic analysis (Braun & Clarke, 2022). This high-level mixed-method framework does not specify detailed methods for its processes. Its hallmark is the reflexive examination of the data by investigators. The roots of thematic analysis (Braun & Clarke, 2022) are in the positivist/postpositivist epistemological perspective (Merriam & Tisdell, 2016) — meaning its purpose is prediction or control, and it assumes that there is an underlying objective, external reality "out there" that can be discovered. The high-level nature of this framework allows various methods to be incorporated into the overall research design - the authors specify no single method. My study compares two specific methods of semantically analyzing and reflexively interpreting the ISERs – automated and manual. I will manually extract the latent themes arising from the semantic themes from both methods and examine them for similarities and differences.

Figure 3

Guiding Frameworks and Methods



Semantic analysis is the automated method that I will employ in this study. This is a member of a family of methods that investigate semantic components based on lexical, grammatical, and pragmatic features (Ignatow & Mihalcea, 2018). In this study, I will use the Latent Dirichlet Allocation (LDA) algorithm (Blei & Mcauliffe, 2007; Blei et al., 2003). LDA belongs to a family of machine learning algorithms called probabilistic topic models (Papadimitriou et al., 1998). These statistical models employ word co-occurrences (words appearing closely together in a text) to identify latent topics or themes within a textual document (Ignatow & Mihalcea, 2018; Matsuda et al., 2018; Soysal & Baltaru, 2021). The general idea, for example, is that if the words bass, rod, stream, and lure appear close together in documents, one could interpret this to mean that the documents discuss fishing (Firth, 1957). Again, I will generically refer to this as the automated method or automated topic extraction method. I provide considerably more detail on the specific approaches I use in the methods section.

The LDA algorithm generates clusters of co-occurring semantic terms. However, the overall topic these semantic terms refer to is treated as a hidden variable representing latent dimensions in the data, and the model cannot offer any canonical topic descriptions (Ramage et al., 2009). Consequently, the researcher applies an interpretive lens to label and characterize the 'aboutness' of the semantic topic, transforming it into a latent topic. This class of models has seen growing implementation in social science research over the past several years (Buenano-

Fernandez et al., 2020; Gil et al., 2021; Ramesh et al., 2014).

Content analysis (Merriam & Tisdell, 2016) is the manual method I will employ in this study. It is exceptionally well documented in the educational research literature - many studies cited in the previous section of my literature review used content analysis. It is a comprehensive methodology — which helps account for its popularity. Krippendorff (2018) describes content analysis in his introductory textbook:

> Essentially, content analysis is “an unobtrusive technique that allows researchers to analyze relatively unstructured data in view of the meanings, symbolic qualities, and expressive contents they have and of the communicative roles they play in the lives of the data's sources (Krippendorff, 2018, p. 49).

In this study, I will use a definition similar to Berger and Lune (2007) and define content analysis as a qualitative method in which the researcher manually extracts topics, themes, and patterns within and across texts. I further describe the techniques I will use to perform the manual qualitative content analysis in the methods section.

Referring to the automated and manual techniques by their family names (i.e., semantic analysis or content analysis) becomes cumbersome to keep track of by the reader. Consequently, from now on, I will refer to the methods in this study as either “automated” or “manual” methods to facilitate reader understanding and comprehension.

While manual content analysis methods have been used extensively in the research literature, several researchers have used research designs similar to those used in this study to extract identity themes from educational institutions using machine learning methods. Perez-Encinas et al. (2021) utilized an LDA model to examine 73,715 college student reviews on a social media platform about their experiences at various study-abroad colleges. The researchers applied domain knowledge to the 20 semantic topics the LDA algorithm produced and used that domain knowledge to identify latent topics. They determined a different constellation of latent topics between students interested in degree-seeking institutions (looking for a degree) and students interested in credit mobility (looking for credit applicable to their home school) institutions. Furthermore, this difference can be used to distinguish between the two groups of students.

Sun et al. (2019) used the corpus of accreditation reports from Washington State K12 schools to identify school improvement strategies. The LDA algorithm was applied to 623 observations from school improvement planning and implementation reports to identify novel improvement strategies. Fifteen school improvement strategies were efficiently extracted from the report texts, which outline several aspects of policies governing school reform efforts. In summary, probabilistic topic models, specifically LDA, are well-studied and well-documented in the educational research literature.

One attraction to these models is that they allow researchers to analyze enormous amounts of contextualized data. Algorithms such as LDA can ingest and process thousands of documents, providing researchers with a new tool to view the literature, a macroscope (Graham et al., 2016, p. 1). Chen et al. (2020) examined a large corpus and employed a novel method of topic model validation. The study used Structural Topic Modeling (STM) — a refinement of LDA — to examine all 3963 articles in Computers and Education published between 1976 and 2018 to identify research trends and opportunities in education. This is noteworthy because the STM latent topics were validated using topics generated by other machine learning methods — LDA and LSA (Latent Semantic Analysis).

Yin and Yuan (2022) conducted similar research, analyzing 3772 abstracts from papers on blended learning published between 2003 and 2021. The big data these researchers used would be difficult, if not impossible, to analyze using strictly manual methods.

A growing number of studies have examined the impact of manual vs. machine-learning textual analysis methods on research findings. Many studies comparing automated and manual approaches involve datasets with easy-to-determine, unambiguous labels to facilitate easy calculation of algorithm performance metrics. Whether these datasets involve sentiment analysis (Borromeo & Toyama, 2015; Kirilenko et al., 2018), medical diagnosis (Rosenberg et al., 1990), or text clustering (Muller et al., 2022), these studies lack the open-ended nature of the documents used in this study. Nonetheless, they offer some meaningful insights that this study builds on.

For instance, early work by Rosenberg (Rosenberg et al., 1990) compared two automated and one manual content analysis method on medical interviews to classify diagnoses. They determined that computerized systems generated more accurate outcomes, speculating that the computer can stay focused longer than human researchers. De Graff and van Der Vossen (2013) explicitly compared the performance of automated versus manual content analysis methods on a large corpus of press releases. They determined: (a) that automated methods are more efficient, but not for small samples; (b) that automated methods are not demonstrably more objective because these methods still require human support; (c) automated methods can still add value to research by offering an independent method of validation, and may be less biased than humans and, therefore, more useful for triangulating results.

Other researchers compared automated versus manual content analysis methods for text clustering. Muller et al. (2022) studied automated topical clustering of journal articles examining the effect of secure institutions for children and youth with behavioral problems. Their aim was the automatic creation of useable thematic categories for topical clustering. They reported that the automated methods failed to find two researcher-discovered clusters but found one cluster the researchers missed.

Some scholars have argued that automated forms of content analysis are only useful for semantic analysis and cannot acquire the more nuanced latent topics present in the analyzed texts (Conway, 2006; Linderman, 2001; Nacos et al., 1991). As Simon (2001, p. 87) explains, “The chief disadvantage is that the computer is simply unable to understand human language in all its richness, complexity, and subtlety as can a human coder.” Sjøvaag and Stavelin added, “human labor is still considered superior for the coding of latent content” (Sjøvaag & Stavelin, 2012, p. 219). Boyd and Crawford (2012, p. 671) conclude, “Context is hard to interpret at scale and even harder to maintain when data are reduced to fit into a model. Managing context in light of Big Data will be an ongoing challenge”.

Automated methods appear to have a slight performance advantage compared to manual context analysis methods for classification tasks. However, the picture changes when a less structured task involves extracting latent topics from a text. In these situations, machine learning algorithms may not perform as well — possibly because extracting latent themes relies on human sensemaking processes.

Automated and manual context analysis methods both have limitations. This may be one reason why hybrid techniques using automated and manual methods are common in the social sciences (Gándara & Daenekindt, 2022). Researchers agree that both automated and manual approaches are synergistic when used together (Lewis et al., 2013; Sjøvaag et al., 2012; Sjøvaag & Stavelin, 2012). “Computational power, ensures quality, precision and scale in registering platform-specific elements … while tried and tested practices of the news content analysis ensure assessment of thematic categorization” (Sjøvaag et al., 2012, p. 93).

Lewis et al. (2013) argues that hybrid automated and manual methods are the only way to deal with data at scale. One way this can be accomplished is by using automated methods to filter and reduce big data sets — freeing humans to interpret the results. Furthermore, humans are very much needed to add contextual dimensions to semantic analysis. A possible approach employing this idea may be using computerized methods to add additional cues for human researchers — thereby improving interrater reliability.

> For example, in traditional content analysis, researchers may give coders a set of keywords as indicators for a given category; a hybrid approach could automatically perform a search for such keywords within a given text and subsequently suggest the

> respective category to a human coder, who would then either confirm or revise the suggested assessment. (Lewis et al., 2013, p. 48)

There are limitations to both automated and manual content analysis methods. Similar to the method used in the study, hybrid methods have been implemented to take advantage of synergies arising from a fusion of the two approaches. Furthermore, this suggests that the problems this study's research questions are designed to address are well-known but not well-solved. Finally, the body of knowledge regarding the hybrid methods discussed here, coupled with the absence of similar literature in educational research, suggests a need exists to introduce the well-known methods and concepts utilized in this study into the broader field of scholarly educational research.

Study Contributions

There are several critical gaps in the literature. First, using automated and manual methods, researchers using mission and vision statements have isolated latent topics from specific colleges. However, whether these two methods extract identical or similar latent topics is still being determined. Second, while mission and vision statements are widely used in research, their small size may generate limited variability in identifiable themes. Accreditation reports, which have received little attention from researchers, might be a richer data source. Third, few papers have explicitly extracted identity themes related to colleges.

Furthermore, it is still being determined whether the latent themes extracted from the literature reached the level of identity themes of the organization. Addressing the literature gaps is essential because it will give administrators more information to use and guide their colleges. Finally, the methods used in this study have been explored in other disciplines but need to be more well-known in educational research.

My study uses a richer source material to identify college identity elements than many previous studies. Also, it contributes methodologically by comparing topics extracted by a machine learning method to topics identified using a well-known method in qualitative research. This method will improve our understanding of the impact of each method on the results it generates. This could have broad applicability to researchers and administrators alike.

CHAPTER THREE: RESEARCH DESIGN

"You shall know a word by the company it keeps."
- John Rupert Firth (Firth, 1957)

Introduction and Overview

This study utilizes a mixed-method research design to ascertain elements of identity for community colleges. I attempted to capture college identities using text from college accreditation reports. Many of these reports contained mission, vision, value statements, and discussions describing how college operations align with accreditation standards. I coded sections of these reports with two methods - one using automated machine learning and the other manually human-driven. Finally, I compared the emerging themes for similarity and dissimilarity. This was a mixed-method study because I employed manual and automated methods in the semantic coding phase. In this section, I first describe the study sample and discuss the data collection process. Then, I discuss general analysis procedures and highlight the procedural differences required for each research question. I have organized the data analysis section by research question because that determines which method I will employ.

Sources of Data

Institutional Self Evaluation Reports (ISERs) were ideal data sources for this study. These documents are part of the collegial accreditation process the United States uses to ensure quality control in higher education institutions. College accreditation is not automatic; colleges apply with various accrediting agencies for recognition. The application process is complex but revolves around the college's compliance with different published standards promulgated by the accrediting agency. Accreditation agencies also have a formal recognition process as well (after successfully completing the application process, the U.S. Department of Education acknowledges the applicant as a "nationally recognized" accrediting agency (Education, 2022).

Accreditation is required for any school receiving federal funds or grants (Education, 2022). Federal law requires accrediting agencies to promulgate accreditation standards for colleges to follow. Noncompliance with these standards could result in the college losing its accreditation (California State Auditor, 2013). This could be disastrous for college enrollment.

The Accrediting Commission for Community and Junior Colleges (ACCJC) is the California Community Colleges (CCC) accrediting body and sets accreditation rules. Every seven years, an individual college must create an in-depth self-evaluation report (the ISER) measuring its performance against the published ACCJC standards (Accrediting Commission for Community and Junior Colleges, 2022). This report is forwarded to the accrediting agency at the beginning of an accreditation cycle. The accrediting agency reviews this report and sends a team to verify the accuracy of the ISER's content. The "visiting team" also produces a report it submits to the ACCJC. The accreditor uses this report and the ISER to determine the college's accreditation status for the next seven years.

The ISER describes the activities the college is actively undertaking to comply with accreditation standards. It is not an aspirational or idealized document and describes the day-to-day reality the college experiences. There is no "right way" to meet the standards, and colleges can implement any policies or procedures they choose to do so. Therefore, it is possible to enumerate the topics important to any specific college by examining the textual content of the ISER. Identity leads to policies and practices, which appear in the ISER as evidence of meeting the specified standards.

Manually analyzing an entire ISER is a nontrivial task. These documents comprise 20 - 25 sections, including the 14 sections enumerating the individual standards. An ISER can run between 200 and 400 pages with text, images, and tables, and colleges spend hundreds of person-hours writing them. There is no technical reason why the entire ISER cannot be analyzed to ascertain a college's organizational identity. However, I studied one specific section to zero in on the efforts colleges consider most central to their identity. There are standards in the ISER that are unlikely to make meaningful contributions to efforts to uncover key elements of organizational identity, such as Standard IIIC: Technology Resources (which addresses computers and other educational technology). Conversely, some standards speak directly to the ideals of the college, such as Standard IA: Mission (which directly addresses the college's mission and values). This project has analyzed text only from Standard IA and excluded the other parts of the ISER.

Data Collection Procedures

Random selection from a list of California Community Colleges published on the state chancellor's office website (California Community College Chancellor's Office, 2023) was used to identify 25 colleges from the 116 colleges in the California Community College system. A random number generator was used to select (without replacement) the colleges from the Chancellor's Office list. Calbright College was excluded from this list because it is an entirely online college — making it very different from the other colleges in this study. This study did not have human subjects, and IRB approval was not required. A list of the college ISERs used is provided in Appendix A: List of Colleges in Sample.

The figures and tables displayed in this section (Research Design) were created using pilot data from 6 randomly selected California Community College ISERs. Two of these ISERs (American River College and College of the Canyons) were also randomly selected for the 25 colleges for the main study. Data in the remainder of this document (i.e., in the Results section) came from the larger study dataset. The two datasets used exactly the same process in creating the LDA model.

This study data analyzed 25 ISERs publicly accessible from their respective college websites before June 30, 2023. The ISER dates ranged from 2017 to 2023. I went to the randomly selected college websites, found the accreditation page, and downloaded .pdf versions of each college's ISER from the last accreditation cycle. These were converted to text files by opening the pdf in Adobe Acrobat (thus confirming the integrity of the downloaded copy) and manually exporting the ISER to a .txt file. This format is more accessible to the computer than the .pdf format. All elements of the ISER in the text file were manually removed except Standard IA. All text files were stored in a common folder, and the manual and automated methods used the same .txt file. These methods are described further below.

Data Analysis

This study utilized Thematic Analysis (TA) (Braun & Clarke, 2022) as an overarching framework for data analysis. I now discuss my approach to making meaning (coding) and the levels of analysis used to code the meaning in this study. An inductive coding approach was appropriate because the research questions focus on the experiences, perceptions, and meanings of the "participants" in the study (Braun & Clarke, 2022). Here, this constitutes looking for meaning in colleges' reported efforts and experiences as reported in their ISERs.

Thematic Analysis is the consistent, overarching framework that binds together several methodologies used in the study. Braun and Clark (2022) describe it as a "method for developing, analyzing, and interpreting patterns across a qualitative data set, which involves systematic processes of data coding to develop themes – themes are [the] ultimate analytic purpose" (Braun & Clarke, 2022, p. 3).

TA does not have a predetermined theoretical framework, making it easy to employ various methodologies. It demanded careful consideration from the researcher and reflection on the different elements of the research design (Trainor & Bundon, 2021). As noted in the literature review, it was an ideal framework for this project because it is highly flexible and easily applied to make sense of codes derived from qualitative textual analysis or machine learning methods. Below, I outline my basic approach to thematic analysis, which combines procedures recommended by Braun and Clark (2022) and Kuckartz (2019) to apply both manual coding and automated methods. Table 2 outlines the general process. The activities in Phases 1, 2, and 3 differ by the research question, while Phases 4, 5, and 6 are constant throughout this study. This results from developing two groups of themes from the ISERs - one group used semantic codes generated by humans, and one drew from semantic codes generated by automated methods.

Table 2

Thematic Analysis Stages

Phase	Stage
1	Intensively Reading | Precoding
2	Semantic Coding
3	Latent Coding
4	Develop & Review Themes
5	Naming Second-Level Themes
6	Present Results

Note: (Braun & Clarke, 2022, p. 6; Kuckartz, 2019)

The objectives of the phases are as follows: Phase 1: The "Intensively Reading | Precoding" phase consisted of an initial familiarization read of the data (qualitative method) (Braun & Clarke, 2022, p. 6; Kuckartz, 2019, p. 188) or cleaning and organizing of the data for future processing (machine learning method) (Mertz, 2021). The "Semantic Coding" phase (2) created the codes from the data. Semantic codes (Oxford Languages, 2023) stuck closely to the language of the document under study. The initial set of codes extracted under each method used the language drawn from each ISER, as described below.

Phases 3-6 involved taking the semantic codes and making sense of them more systemically. Phase 3 aggregates semantic codes into larger themes. Phase 4 is a general check to ensure the larger themes make sense. The last step in the process, phase 6, presents the themes sensibly.

The "Latent Coding" phase (3) abstracts the semantic codes into larger concepts. These composite latent codes (Kuckartz, 2019, p. 183) attempt to capture a deeper, more conceptual meaning. The implicit assumption was that latent codes are better descriptions of organizational identity because they are more conceptual and abstract. This study assumed that organizational identity was not easily described in a word or two, but rather, identity was comprised of an amalgam of complex concepts often not directly articulated.

The "Develop and Review Themes" phase (4) was a re-engagement with the data to identify any better abstractions of the codes. Developing rich themes was the goal of this phase and served as a sanity check (‘Do these themes even make sense?’). The “Naming of Second Level Themes” phase (5) was about creating a theme definition - enumerating the central organizing concept of each theme as well as highlighting any particularly illuminating manifestations. The “Presenting Results” phase (6) was a recognition of the fluidity of qualitative research and the possibility of the “themes [shifting] until the final manuscript was published” (Trainor & Bundon, 2021, p. 14).

One key question was the order in which I conducted the parallel manual coding and automated coding analysis. One concern was that if I conducted the manual analysis first, my familiarity with the underlying text could contaminate my latent coding of the automated semantic codes. While my review of the automated codes could have theoretically contaminated my manual codes, this was less of a threat since the automated codes provided less context and immersion of the researcher in the text. Therefore, this project started with work on the automated methods used for research question 1 to minimize my exposure to the ISER content. There was minimal human involvement in producing the semantic topics for the automated method beyond that required to set preprocessing parameters, so there was no need for me to thoroughly read the text before computer processing began. Consequently, the researcher’s (lack of) familiarity with the underlying text did not contaminate the creation of latent topics for this method.

This study compared two methods to generate code labels and themes, resulting in significant procedural differences in the first three phases. Below, I discuss each research question and detail the analytic process.

Research Question 1

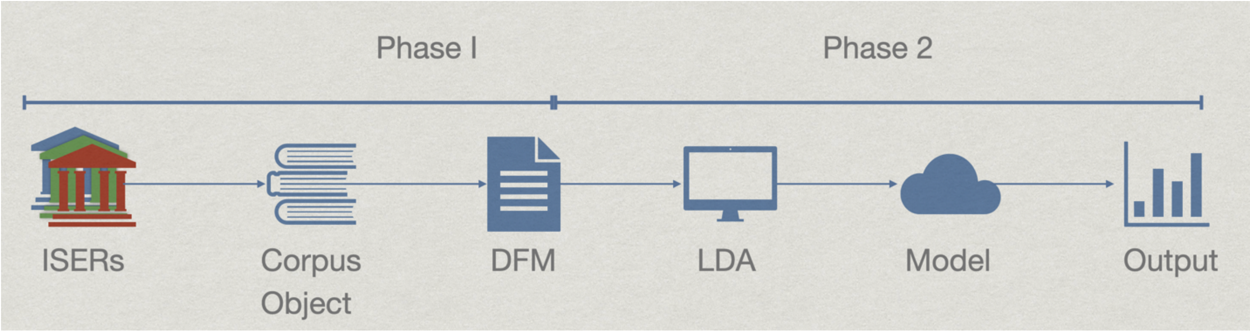
RQ 1. What are the main identity themes of CA community colleges as determined by automated topic extraction analysis of ISER Standard IA?

Answering RQ1 utilized the process shown in Figure 4. College ISERs enter the process on the left and comprise the corpus object. Precoding and preprocessing occur before this step. The

Corpus Object is created when the data is loaded into the computer. Once the data is ready for processing, the computer will generate a Document Feature Matrix (DFM) as the primary data structure for processing (sample below in Table 3). The LDA Algorithm uses the DFM to create a model of the dataset. This model clusters the terms together into topics. Finally, the predictions are displayed as output.

Figure 4

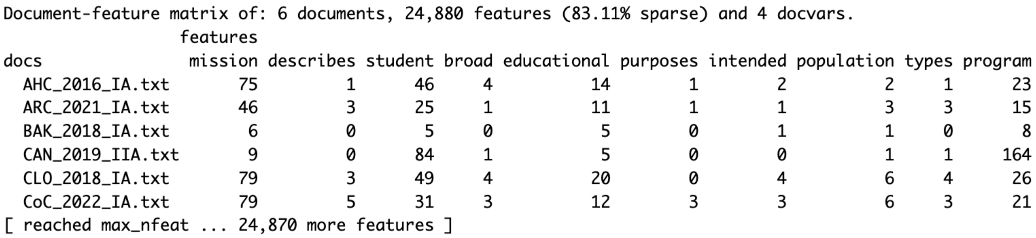
Process of Automated Topic Extraction



Once the terms were in a format the computer could process, a data structure was created in memory for the computer to hold the terms while processing them. Table 3 is an example that I created to pilot these methods using test data from six randomly selected ISERs. Throughout this section, I present examples from this pilot sample to illustrate the methods used. Figures and Tables produced by the actual experimental data are provided and interpreted in the Results section.

Table 3

Sample Document Feature Matrix (DFM)



```
Document-feature matrix of: 6 documents, 24,880 features (83.11% sparse) and 4 docvars.
                  features
docs               mission describes student broad educational purposes intended population types program
  AHC_2016_IA.txt       75         1      46     4          14        1        2          2     1      23
  ARC_2021_IA.txt       46         3      25     1          11        1        1          3     3      15
  BAK_2018_IA.txt        6         0       5     0           5        0        1          1     0       8
  CAN_2019_IIA.txt       9         0      84     1           5        0        0          1     1     164
  CLO_2018_IA.txt       79         3      49     4          20        0        4          6     4      26
  CoC_2022_IA.txt       79         5      31     3          12        3        3          6     3      21
[ reached max_nfeat ... 24,870 more features ]
```

A Document-Feature Matrix (DFM) table stores the data before the LDA algorithm begins processing. Each row starts with the document name, and each subsequent column holds a count of each term ("feature") in the corpus. The documents column shows the file names of the six pilot ISERs. This small DFM holds 24,880 terms, and it is easy to imagine how large this number could grow with even a modest dataset. A second interesting item is that the DFM consists of mostly blank cells, 83.11% of all cells. This is because many words appear in only one or two documents, and the remainder of the cells are blank. Unfortunately, the computer will still need to examine all the cells in the DFM, which is one reason why text processing is computer-resource intensive.

Once this phase was completed, the text documents were difficult for humans to read, and the DFM was not designed to make sense for humans – the DFM is a simpler corpus created by removing extemporaneous and redundant data for the machine learning LDA algorithm.

Phase 1 – Precoding.

The dataset was preprocessed in the precoding phase for machine learning methods so the computer algorithm could use it (Mertz, 2021). The preprocessing phase converted raw human-readable textual data into something the computer can analyze. This is the beginning of Phase I in Figure 4. The LDA algorithm expected the removal of symbols with little or no meaning to the topic modeling process. For instance:

- Punctuation was removed.
- Numbers (1, 2, 3, etc.) were removed.
- Symbols (&, #, etc.) were removed.
- Uppercase letters were converted to lowercase.
- Stopwords were removed. A stop word is a common term that adds no information to a sentence. Examples in English are: "is," "the," "a," and "an."
- The titles for the various standards were manually extracted. For example, "iv.d" or "ii.b.4".

- Words were tokenized. The computer converts the text into individual words as units of analysis (tokens). This study converts text into individual word tokens by detecting white space and using it as a token boundary. I also introduced an additional level of granularity by including more adjacent words as our unit of analysis. In this case, I used bigrams (two-word tokens).
- Minimum frequency words were removed. Words that do not occur often enough do not carry any information. The term "Allen" likely only appears in one document (the Allen Hancock College ISER from the test dataset, for example) and is unlikely to be related to terms outside that particular college.
- Blank row check. The LDA algorithm does not perform well when there is a blank row; this step checks to ensure there are not any in the documents.
- A dictionary stemmer reduced many (but not all) words to a common term. The words "students," "student's," "Students," and "Student's" were all reduced to the common root of "student," for example. Otherwise, the computer would not recognize that these terms all belong to the same concept of "student."
- Composite entity construction. The various colleges in the ISERs often refer to themselves by name. This can create a problem in the corpus because LDA focuses on groups of related terms. If this issue is not addressed, the corpus will fill up with associations such as "Irvine Valley College" + "strategy," "Saddleback College" + "strategy," and "Folsom Lake College" + "strategy." The problem is that each of these term pairs will be considered a separate LDA topic, diluting the apparent size of the topic "strategy" in the corpus. The solution was to replace all specific college name references with a generic composite entity name, "the_college." All three examples above would populate the corpus with "the_college" + "strategy" and thereby increase the frequency of this topic in the corpus.

Once this processing was completed, the ISER texts could be loaded into the computer to create the Corpus Object shown in Figure 4.

Phase 2 - Semantic Coding.

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) was ideal for semantic coding of the dataset because the algorithm is exclusively focused on terms and other terms that co-occurred with them. LDA has a long history of use in text analysis in the social sciences (Ignatow & Mihalcea, 2017). The LDA algorithm was implemented in the R programming language using the topicmodels package (version 0.2-13). The algorithm identified clusters of words occurring together and generated lists of words clustering around each potential theme (called a “topic” in Topic Extraction).

LDA is based on Bayesian, rather than classical, statistics (Blei et al., 2003). Probabilities are viewed as expressions of subjective beliefs and may be updated when new data becomes available. This allows the algorithm to progress from what is assumed about the phenomenon under study (called a prior probability) and extrapolate (using data and the LDA model) to make probabilistic statements about the parameters of the phenomena under study (called a posterior probability) (Lambert, 2018). The LDA algorithm can use researcher-supplied prior probabilities or, if not available, a uniform prior probability. This project used a uniform prior probability from the uniform Dirichlet distribution, where each term was initially assumed to have an equal probability of belonging to each topic (Blei et al., 2003).

A Dirichlet distribution has several attractive properties for topic modeling, including a) the sum of all parameters must equal one, thus ensuring the distribution represents a valid probability distribution, b) it is symmetric and therefore insensitive to the ordering of the parameters and, c) the researcher can control the dispersal of the distribution with the alpha parameter. A larger alpha will concentrate the distribution (and assumes documents possess more topics), while a smaller alpha will disperse the distribution (and assumes the documents are comprised of fewer topics). A higher alpha is often used when the data is expected to have many topics or when more granular topics are desired. A lower alpha is used when the data is expected to have a smaller number of topics or when researchers want more general topics produced by the algorithm (Bishop & Nasrabadi, 2006; Gelman et al., 2013).

Alpha is intertwined with another LDA hyperparameter, *K* – which specifies the number of topics the LDA model should generate. Hyperparameters control the model's behavior and are set before training the LDA model. *K* can be tuned to generate a large or small number of topics, impacting topic granularity, like alpha (Silge & Robinson, 2017). Researchers have identified significant complexities in choosing the proper number of topics for a given corups:

> In general, users prefer models with larger numbers of topics because such models have greater resolution and are able to support finer-grained distinctions. Unfortunately, we have observed that there is a strong relationship between the size of the number of topics and the probability of the topics being judged as nonsensical by domain experts (Mimno et al., 2011, p. 262).

Too large a *K* value results in topics so precise humans have difficulty assigning a label to – a situation I call topical decoherence. However, researchers need to tread carefully as too small of a *K* value generates "vacuous" topics that are overly general (AlSumait et al., 2009).

Unfortunately, there is little guidance in setting the optimum number of topics to look for (Wallach et al., 2009). LDA is a generative Bayesian model and does not produce test statistics like those found in classical hypothesis testing (such as a p-value) (Lambert, 2018). This forces researchers to use one of several potential ways to evaluate the quality of the predictive model instead.

LDA's roots began in the psychology community, and early evaluation of LDA topic models was focused on approximating human performance (Chang et al., 2009b). There was little effort spent attempting to understand the accuracy of the models' latent space; instead, efforts were focused on producing valuable and applicable output. The goodness of fit was determined by how subjectively useful humans found the output (Silge & Robinson, 2017).

Several studies exist evaluating model quality in the literature, but progress has been slow with Ramage et al. (2009) asserting “characterizing topics is hard.” Several studies have explored topic evaluation by comparing model output with some gold standard, either document metadata or domain expert-created labels (AlSumait et al., 2009; Chuang et al., 2013; Mimno et al., 2011). Others have used metrics such as perplexity, which captures surprise or novelty when an LDA algorithm is trained on one dataset and then exposed to a hold-out dataset that it has not seen before (Mitchell & Mitchell, 1997; Wallach et al., 2009). Unfortunately, these methods are often unsuitable for topic evaluation either because they focus on corpus (rather than individual topic) level evaluation or because they perform poorly when compared to topics that humans would extract, suggesting that they provide a poor proxy for human judgment.

Overall, assessing the quality of the LDA model output is challenging because of the high dimensionality of the fitted model (Sievert & Shirley, 2014). Consequently, data visualization is considered appropriate for LDA model analysis (Chaney & Blei, 2012; Gardner et al., 2010; Sievert & Shirley, 2014).

Early work selecting the proper value of K using the pilot data suggested that values less than ten would work well with the ISER corpus. Trial and error in creating an LDA model with different K values were used to explore the potential topic space. Figures 4 through 8 illustrate this process using the pilot dataset.

Figure 5

Plot of Topics in 2D Space Using Pilot Data, K=3

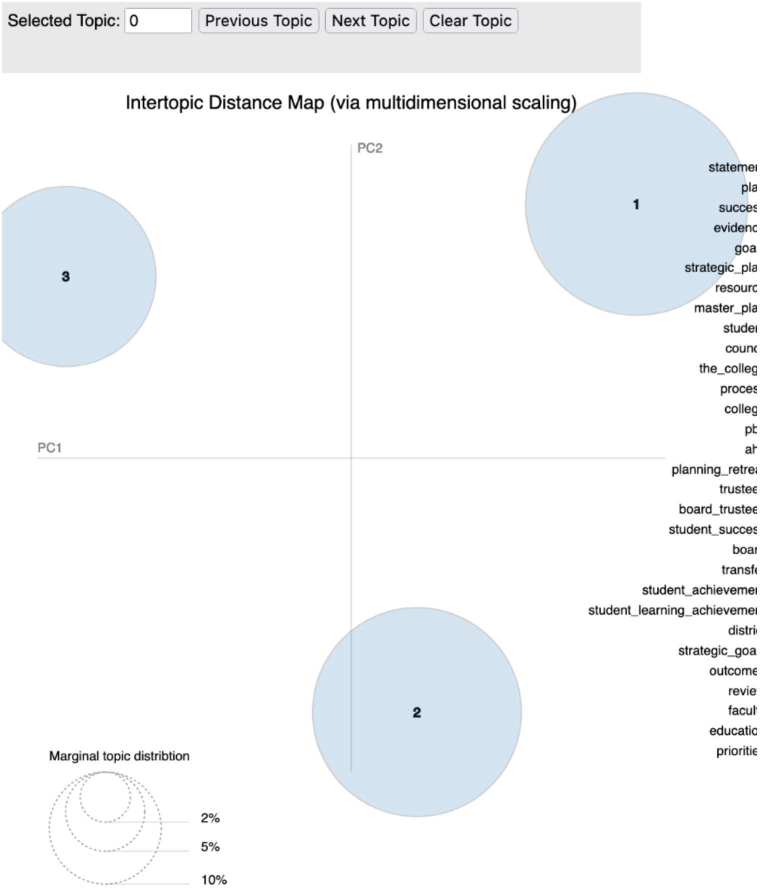
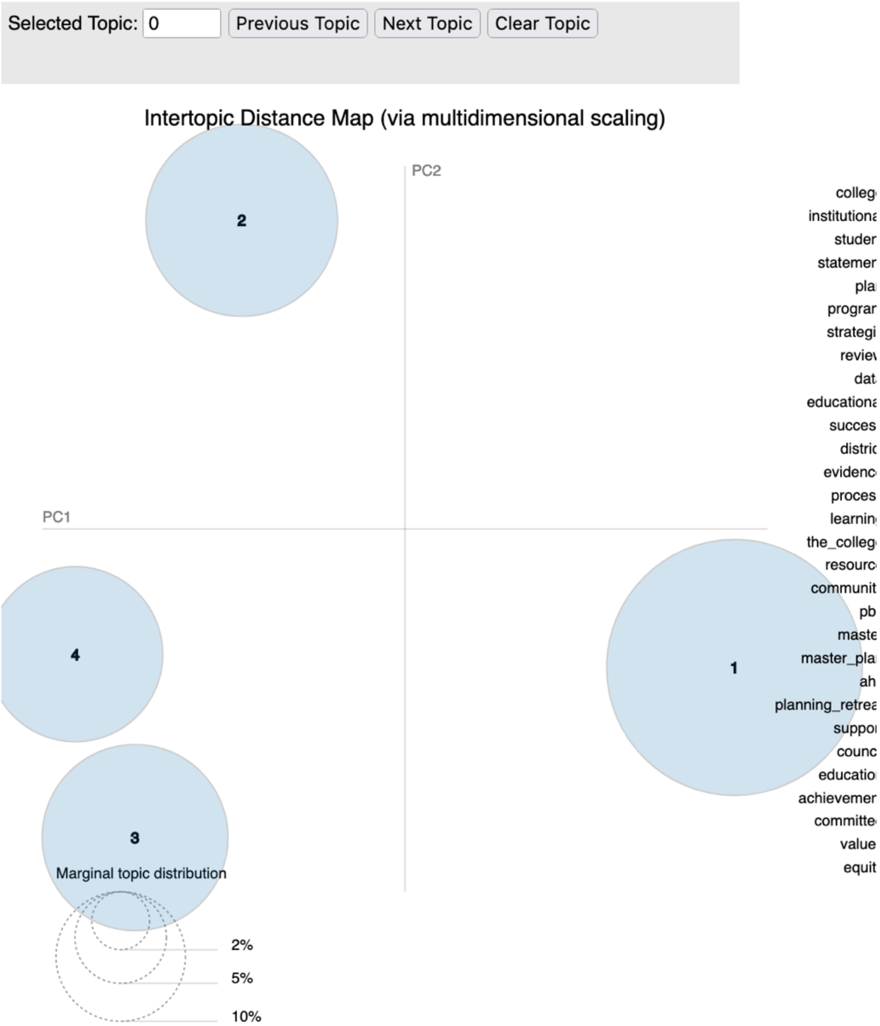


Figure 5 is the left pane from LDAvis (Sievert & Shirley, 2014). This web-based interactive visualization tool provides a high-level overview of the model topics. Additionally, it plots the topics on a two-dimensional space using multidimensional scaling to assess orthogonality. The PC1 and PC2 axes do not have any intrinsic values, although researchers may be able to determine labels for them. The size of the circle indicates how much that topic occupies the corpus. The distance between circles shows how much (if any) overlap topics have. In Figure 5, all three topics have relatively similar importance in the corpus. The blue color indicates that specific topic has not been selected for display on the right side of the pane (terms for that side are partially shown). The topic would be red if it were selected (not shown here).

Figure 5 shows that the three topics are orthogonal, with good separation between topics, but they may also still contain other useful topics. Increasing the K value is the only way to know if that is true. LDAvis cannot visualize fewer than three topics, which is why this process starts with K=3.

Figure 6

Plot of Topics in 2D Space Using Pilot Data, K=4



In Figure 6, the LDA model has split one of the topics. Notice how closely spaced topics number 3 and number 4 are. Using K = 3 or K = 4 would be a judgment call because topics 3 and

4 are so closely related.

Figure 7

Plot of Topics in 2D Space Using Pilot Data, K=5

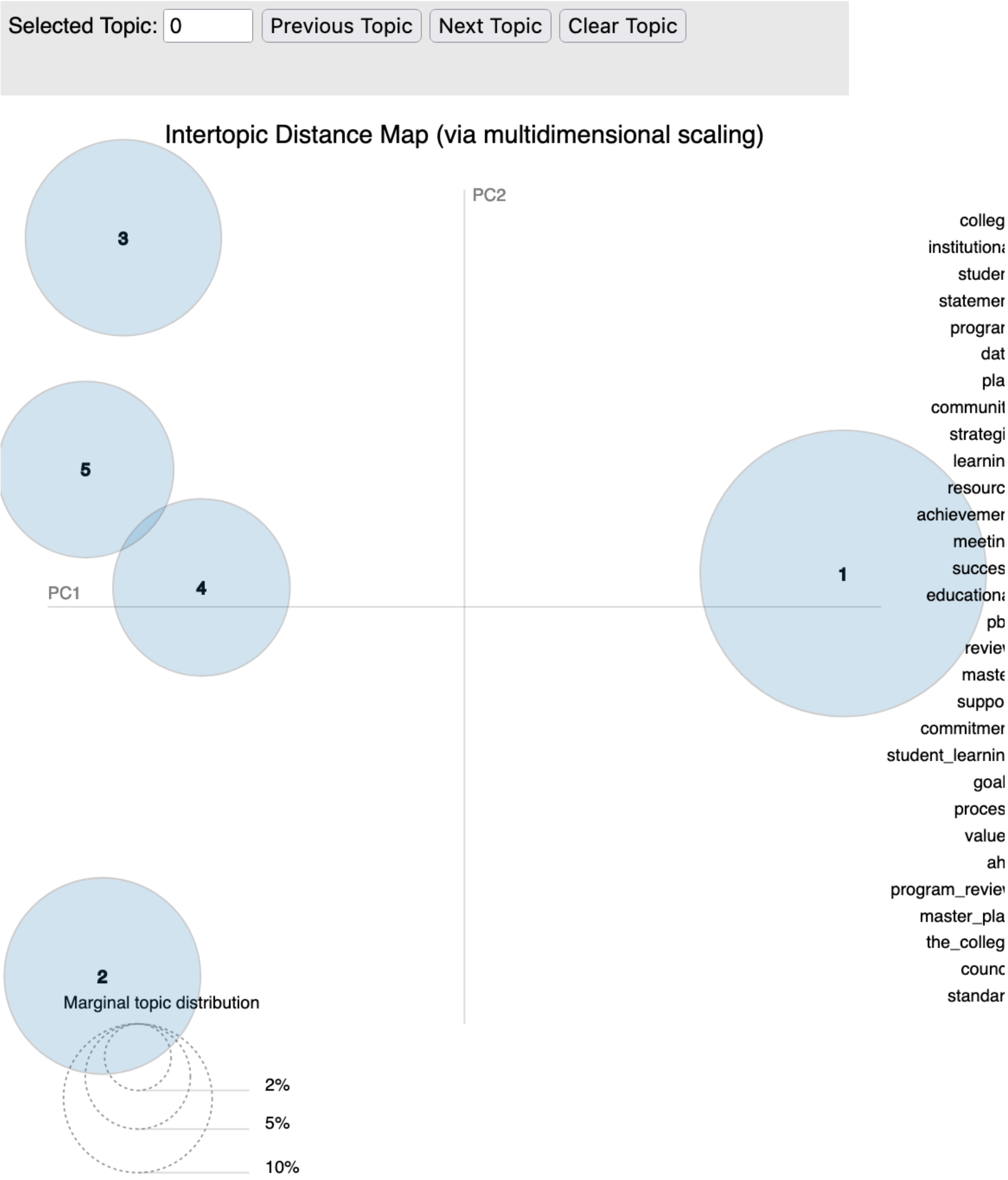


Figure 7 still displays roughly orthogonal topics, but topics 4 and 5 appear to be merging together. Also, notice that the topics have again shifted positions on the Intertopic Distance Map.

Figure 8

Plot of Topics in 2D Space Using Pilot Data, K=6

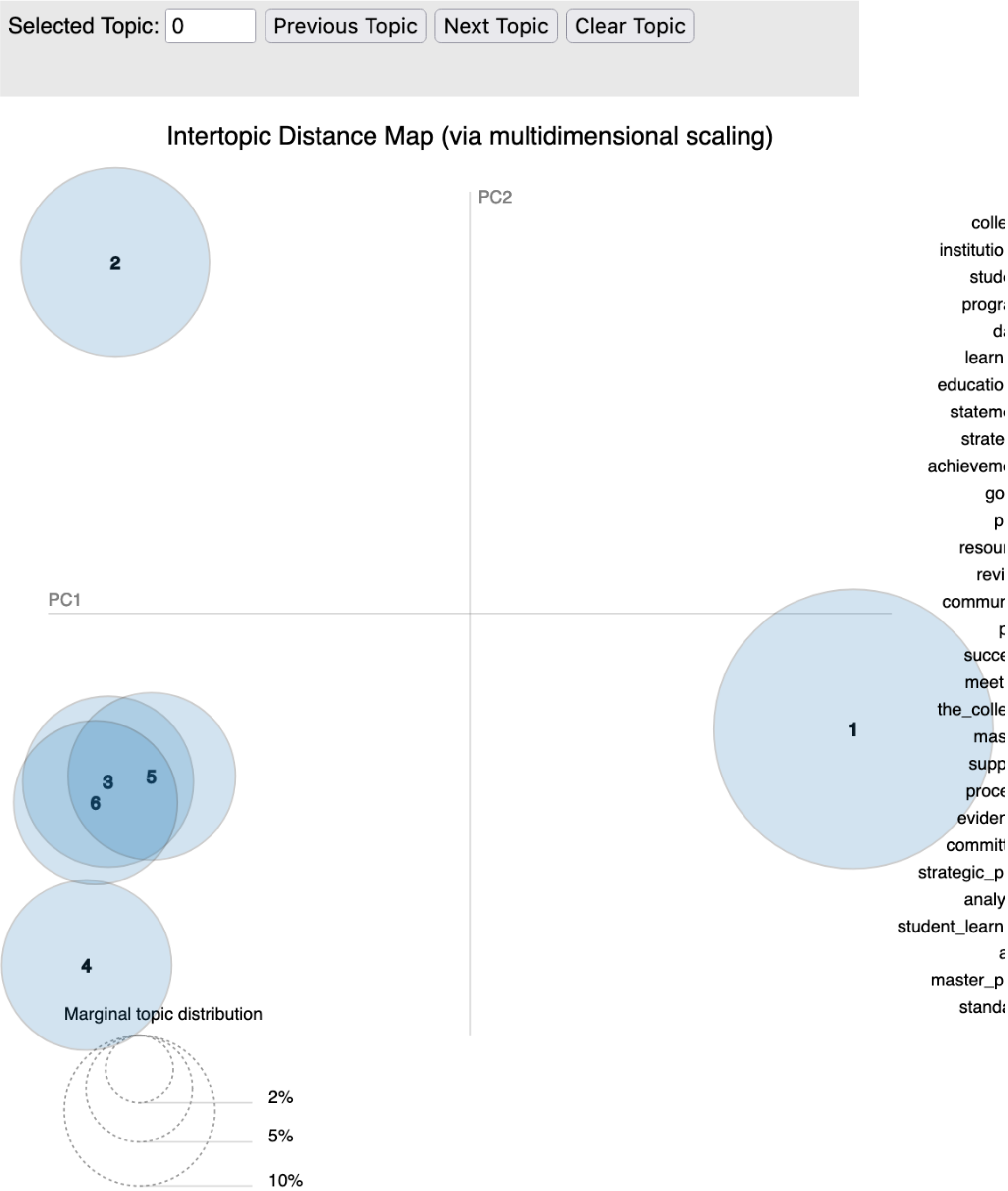
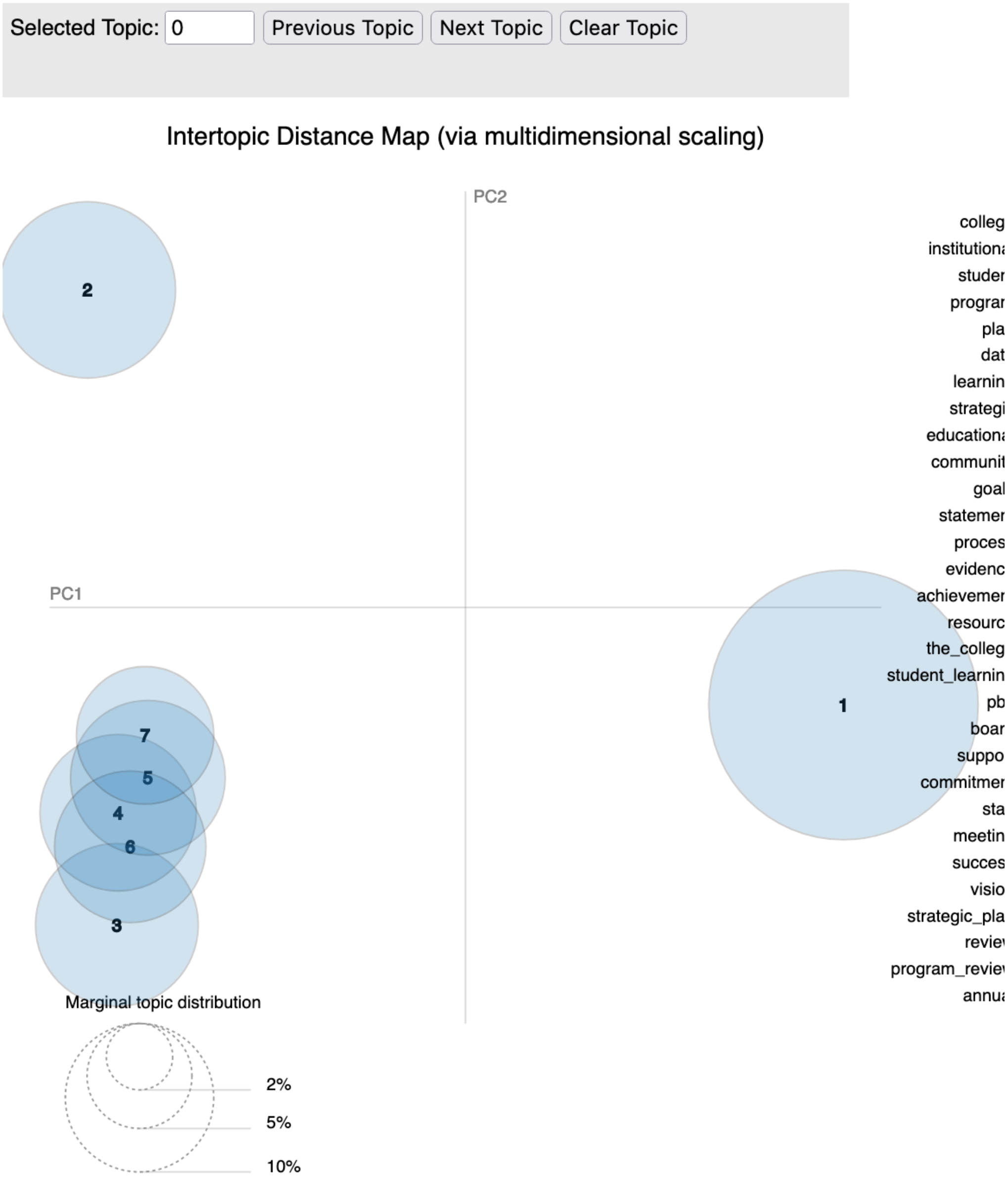


Figure 8 is starting to show topical decoherence. One of the topics from Figure 7 has been broken into topics 3, 5, and 6. Topic 4 is also very closely related to this cluster as well. Notice the significant amount of overlap between this cluster of topics. Also, notice that the location for topic 1 appears relatively stable.

Figure 9

Plot of Topics in 2D Space Using Pilot Data, K=7



More topical decoherence is exhibited in Figure 9. Notice the considerable overlap between the fragmented topics on the lower left side of the pane. This continues the fragmentation observed when K=6 and suggests further increases in K will not be fruitful. Therefore, with the six colleges used in the pilot study, a value of 3 is appropriate for the K parameter because it produced very orthogonal topics.

This trial and error process was also used on the 25 college corpus of the main study. Patterns similar to those discussed in the tables above were observed. As discussed in the results section, it was determined that K=5 maximizes topic orthogonality without the topics decohering. Therefore, K=5 was selected for the main study.

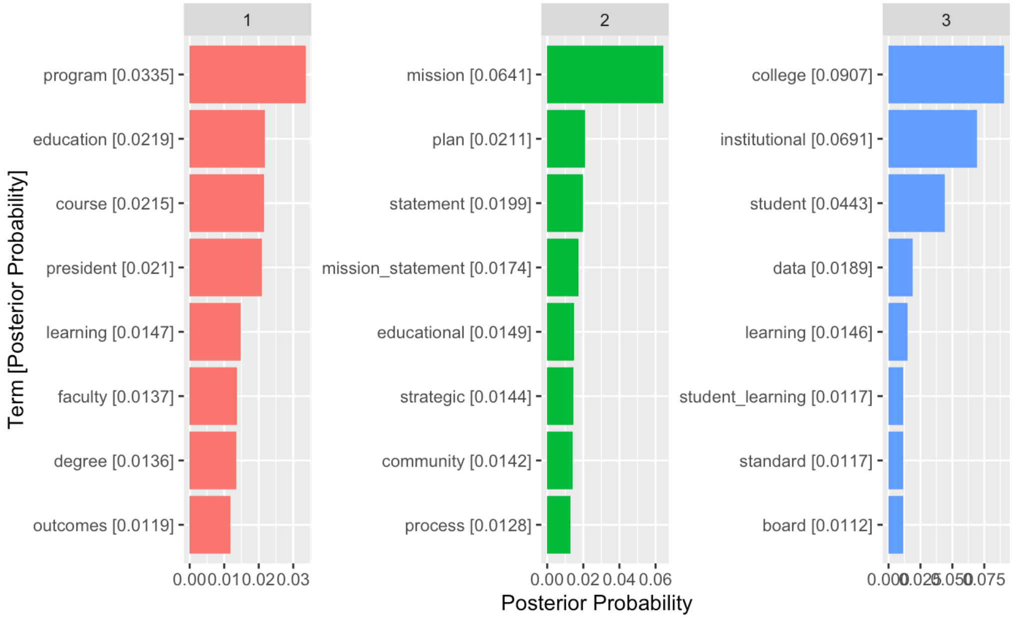
Thus far, the algorithm has identified non-overlapping topics and labeled them "topic 1, 2, etc. " because it has identified semantic relationships between the words comprising the clusters. Note that some of these keywords are provided in the right-hand display of the figures shown above. However, the algorithm can only recognize words as patterns of characters and not interpret their meaning. I provided meaning in the next phase - phase 3, Latent Coding.

Phase 3 - Latent Coding.

I use three key criteria to determine an appropriate list of words within each topic to consider in creating latent topic labels for each topic identified using the process in Phase 2: posterior probabilities, lift, and term prevalence.

Figure 10

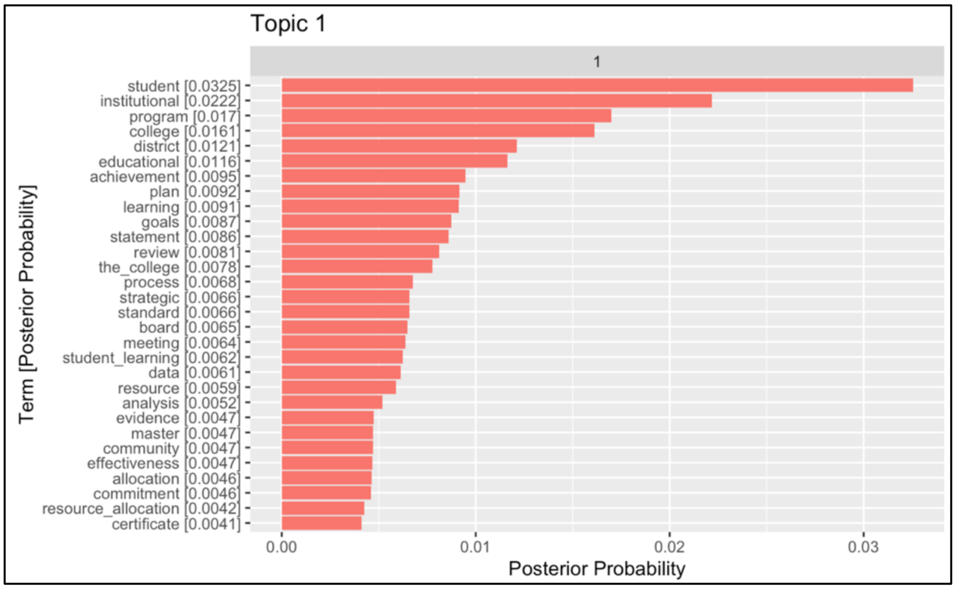
Top 8 Terms by Topic Cluster for Pilot Data (K=3)



Note: Display uses data from a pilot sample of ISERS. The topic numbers in this tool are opposite of those shown because of software differences. This does not impact analysis.

Figure 11

Top Terms for Topic 1 by Posterior Probability



Ranking posterior probabilities (Lambert, 2018) is one way to examine the data, as these probabilities represent the computer’s best guess that particular terms belong to a specific topic.

Figure 11 shows the frequency by the posterior probabilities of the top terms for topic 1 of the experimental corpus (the posterior probabilities for all the topics can be found in Appendix C: Posterior Probabilities by Term and Topic). This provides a sense of a given term's usefulness in deciding the appropriate latent topic. The posterior probability is a confidence estimate that the given term is a member of a given topic. Terms with higher posterior probabilities may be considered more representative of the topic. Terms near the bottom may have posterior probabilities, an order of magnitude smaller than those near the top, contributing much less to the 'aboutness' of the topic. Some of the posterior probabilities can be quite small. Given the distribution of these probabilities in my data, I opt to automatically include terms in consideration for topic labels if the posterior probability is above 0.01.

"Lift" (Taddy, 2012) is the second key metric I use to determine other terms that should be considered in topic labels. Figure 12 provides an enlarged view of the right-hand side display of the pilot data with *K=3*. This provides the list of terms the computer extracted for Topic 1, displayed by relevancy. This image was useful because posterior probabilities were insufficient for determining the significance of terms to the latent topic since any given term may appear in multiple topics — none of which are about that term. In Figure 12**,** the blue bar for each term shows that term's frequency in the overall corpus, and the red bar shows the frequency of the term within this topic. The difference between the two colored bars is called lift, a measure of "aboutness." Lift is the ratio between the probability that a document containing word *w* is also classified as topic *t* to the probability that a document contains word *w* regardless of its topic (Taddy, 2012). The LDAvis display provides users with a visualization of lift:

> By comparing the widths of the red and gray bars for a given term, users can quickly understand whether term is highly relevant to the selected topic because of its lift (a high ratio of red to gray [blue]), or its probability (absolute width of red) (Sievert & Shirley, 2014, p. 68)

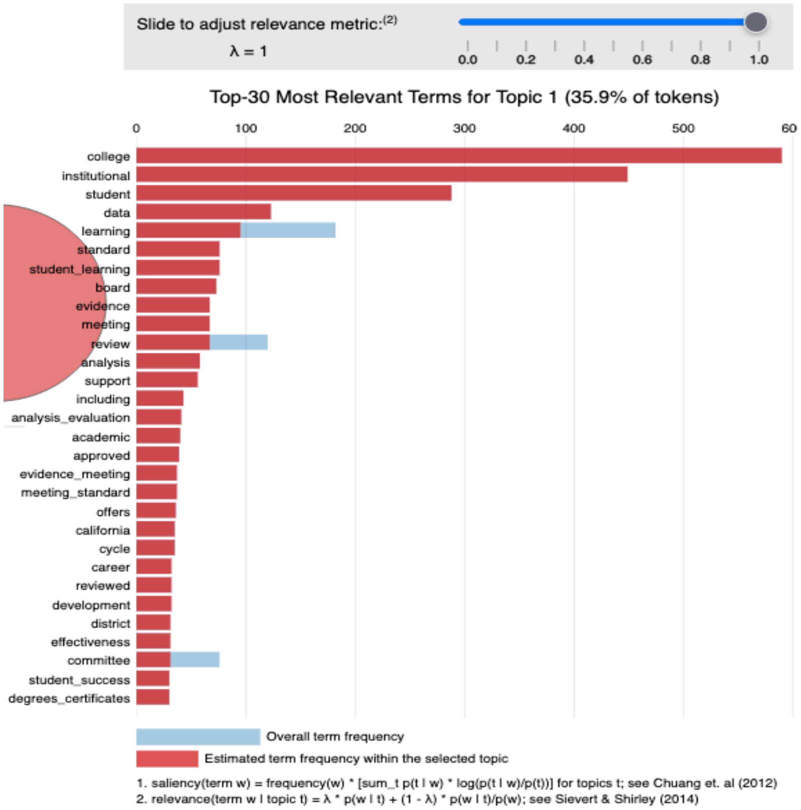
Terms without a blue bar (the word ‘college,’ for example) have a high degree of lift within a given topic because they effectively occur only within that latent topic. By contrast, terms possessing a blue bar (‘learning,’ for example) have a lower degree of lift within a given topic, as they are discussed more broadly outside that topic and appear in other topics (as well as within the current topic). The higher the lift, the more relevant the term is for the selected topic, or the topic is more “about” that specific term.

Additionally, the bars show the frequency of the various terms. For instance, for Topic 1, ‘data’ is far more important to the latent meaning of the topic than ‘committee,’ even though Topic 1 encompasses both terms. While multiple prior researchers have discussed lift (Sievert & Shirley, 2014; Taddy, 2012), they do not offer guidelines around hard and fast values that constitute strong lift. The utility of the metric appears to vary from corpus to corpus. I opt to include terms for consideration in topic labeling if their lift is 100% (i.e., the bar is entirely red) within a given topic area.

The final criterion I use to determine which terms to focus on in labeling a topic area is “term prevalence.” Once lift begins to drop (i.e., a blue bar starts to appear), the exclusivity of the term shrinks, so the task is to determine which topic contained the preponderance of the use of that term. A term that dominated one topic and was present to a minor degree in other topics still significantly contributed to the ‘aboutness’ of that topic – term prevalence. This metric is similar to the exclusivity metric by Bischof and Airoldi (2012) and the relevance metric of Sievert and Shirley (2014), but it is not a numerical score but rather a binary determination. The following process was used to determine whether each term possessed prevalence within a given topic.

Figure 12

Closeup of the Right Side of LDAvis Output for Pilot Data, Topic 1 (K=3)



Slide to adjust relevance metric:[2]

λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t l w) * log(p(t l w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w l topic t) = λ * p(w l t) + (1 - λ) * p(w l t)/p(w); see Sievert & Shirley (2014)

Turning back to Figure 12, the term “review” possesses a blue bar smaller than the red bar, meaning this term is most prevalent in this topic (more than 50% of the term’s mentions are contained in this topic). Contrast that with the term “committee.” That term has a blue bar larger than a red bar and suggests one of two things: a) there is another topic where this term makes a more significant contribution to ‘aboutness,’ or b) the blue bar is spread out over multiple topics, thus making this topic the one where it makes the most significant contribution. Figure 12 cannot answer the question of which situation exists here, and the term must be manually examined. The researcher obtains more detail by clicking on it, which results in a display that scales the topic circles on the left side of the screen by the prevalence of that term in each topic. The researcher identifies the largest circle for each term to determine which topic each term belongs to with respect to term prevalence.

The following codebook in Table 4 was used to prepare the semantic codes' transformation into latent codes. “Item” either refers to individual LDA-generated terms or the various criteria used in assessing distinctiveness. Measurement is the type of data the data assessment method generates. Meaning is the definition of that column in the table. A table using these elements appears in the results section for each LDA topic.

Table 4

LDA Analysis Codebook

Item	Measurement	Meaning
Term	Textual Data	Candidate semantic term identified by the LDA algorithm.
Posterior Probability	Numerical Percentage	The LDA algorithm’s best guess this term belongs to this topic. Range is from 0% to 100%.
Lift	Boolean value	Does this term exhibit high lift (no blue bar in the relevant term bar chart)
Term Prevalence	Boolean value	Does the predominance (>50%) of this terms’ occurrences happen in this latent topic?

Terms that did not have high lift nor predominating term prevalence were largely excluded from consideration in labeling topics unless they clearly contributed by adding nuance to previously identified, stronger terms. Terms with low posterior probabilities were likewise excluded unless they also clearly contributed to previously relevant terms. These less-relevant terms could, therefore, be used primarily to refine or bolster the intuition for topic labels, but only in context with other, more substantial evidence.

Using these criteria, I might assign the latent topic of “Rational Degree Delivery Process” to Topic 1, for example, because the terms “data,” “standard,” “student learning,” “board,” “evidence,” “meeting,” “analysis,” “support” all show a high degree of relevance and lift (contextual information) in Figure 12. The terms ‘student’ and ‘cycle’ both have high degrees of lift, but ‘student’ appears far more often within Topic 1 than ‘cycle’ does — suggesting the latent theme has more to do with ‘student’ than ‘cycle.’

Research Question 2

RQ 2. What are the main identity themes of CA community colleges as revealed by manual topic extraction analysis of ISER Standard IA?

Phase 1 – Precoding.

The Intensive Reading/Precoding phase for the manual process involved in research question 2 involved familiarizing myself with the dataset by reading the entire text of Standard IA individually for each college. My objectives for this phase were: 1) immersing myself in Standard IA of the ISER and developing a deep understanding of my dataset, 2) critically reading the text as data with intimacy and distance, and 3) recording notes about the process. The text from the college's ISER Standard IA was loaded into Dedoose for reading and analysis.

Phase 2 - Semantic Coding.

Semantic coding for this aspect of the project involved *in vivo* coding (Saldaña, 2021) for each concept in each ISER Standard IA. In vivo coding uses the actual words in the data for codes, so it is semantically based by definition. I read the text for "data of analytic interest" (Braun & Clarke, 2022, p. 61; Trainor & Bundon, 2021), which, in this case, are terms that may have been indicative of elements of organizational identity.

I organized codes using Dedoose (Dedoose, 2023). This was an iterative process. In my phase 2 coding, I did not concern myself with the relationship of the codes to the research question, theme cohesion, or even the potential viability of the code. In this phase, I aimed to identify codes - single ideas that can be attached to pieces of text. For instance, MiraCosta College's ISER, in my pilot ISER data, identifies its mission as follows:

> The MiraCosta Community College District mission is to provide superior educational opportunities and student support services to a diverse population of learners with a focus on their success. MiraCosta offers undergraduate degrees, university-transfer courses, career-and-technical education, certificate programs, basic-skills education, and

> lifelong-learning opportunities that strengthen the economic, cultural, social, and educational well-being of the communities it serves. (MiraCosta College, 2016, p. 86).

In my iterative phase 2 coding process, I applied sample in vivo codes to such text (as well as to the other text in standard IA of the ISER), identifying potential OI concepts. In the above text from the pilot dataset, sample codes "educational opportunities," "student supports," and "student success" were extracted.

Phase 3 - Latent Coding.

Now, I will discuss the manual process involved in phase 3 - latent coding. This phase combined semantic codes into candidate themes – overarching ideas (Braun & Clarke, 2022). My goal was to aggregate the codes into more contextually relevant concepts. Manually generated codes were organized in Dedoose to combine semantic codes into latent themes linked by central thematic concepts and concept boundaries. For example, semantic codes of "transfer courses," "certificate programs," and "undergraduate degrees" may be combined into the latent theme "support student degree and credential goals." The resulting latent codes were aggregated and used in phase 4.

I will now discuss the last three phases of my method. The prior three phases needed to be discussed by the research question because the methods employed demanded different processes. However, once the latent coding phase is complete, the data follows the same process regardless of the method used to create it.

Phase 4 -Developing and Reviewing Themes

Phase 4 - Developing and reviewing themes is where I reviewed the viability of my initial clusters and ascertained if any patterns were unaccounted for or not fitting in the topics. I am checking to see if the topics are rich. I wanted the themes to revolve around a central idea and be distinctive. Ideally, these themes should support each other to tell an overall story about the data.

However, suggesting that analysis only occurs during this phase would be misleading. Merriam and Tisdell (2016, p. 195) explain, “collection and analysis should be a *simultaneous* [italics authors] process in qualitative research.” After reviewing every ISER, my field notes documented my observer comments, hypotheses, hunches, and conclusions. I continued this process until I reached the point of saturation — a state where continued analysis yields no new information or insights (Merriam & Tisdell, 2016, p. 198).

Phase 5 - Naming Themes

Phase 5 - Naming Themes is about finalizing my latent themes. Theme labels were decided upon. The boundaries between concepts were clarified. For instance, semantic codes of “Ethics and Integrity …,” “… dedicated to behaving ethically …,” “… we value … integrity,” and “… integrity as the foundation for all we do” may be combined into a latent theme of “Integrity valued.” The latent themes in Table 5 represent OI concepts emerging from synthesizing the semantic themes. The terms in quotations in the first column are the shorthand labels for the latent theme. The central concept discusses what makes this concept different and distinctive from the other latent themes. The exemplars are direct quotes from the various College ISERs, including relevant semantic codes.

The process of constructing more complex concepts from simpler semantic codes slightly differed from the manual themes. The in vivo coding resulted in a more extensive spread of codes. For example, under the semantic code of “Equity_Valued,” the one-word in vivo code of “equity” was common. However, phrases such as “equity-focused environment” and “equity and social justice in student outcomes” were also present. This would not have lent itself to the relatively straightforward process presented in Table 5 because the concepts were semantically complicated.

Table 5

Theme Definition for LDA Extracted Topics, Pilot ISER Data

Topic Number	Latent Theme	Central Concept/Scope/Boundaries of Theme	Exemplars/ Direct Quotes
1	Rational Degree Delivery Process	The institution uses a uniform, data-driven, evidence-based process to deliver student learning.	"data," "standard," "student_learning," "board," "evidence," "meeting," "analysis," "support"
2	Mission Preeminence	The Mission and Vision statements drive subsequent college processes. The statements are the hub upon which the spokes of the other college processes and activities revolve.	"mission," "mission_statement," "plan," "strategic"
3	Traditional College Education Process	The college is oriented around programs of study, comprising of courses, focused on delivering degrees.	"program," "education," "course," "president," "faculty," "degree" "outcomes"

However, once the semantic codes were collected, the topic labels were generated by collapsing related themes. Creating the topic of "Principles of Social Justice" involved aggregating the semantic codes of "Equity_Valued," "Diversity_Valued," and "Mutual_Respect," for instance. The process articulated above and Table 5 were the foundations for subsequent analysis to address research questions 3 and 4.

Potential Threats to Methods

Discussing potential confounds before moving on to the final section is instructive. As noted above, this study's design may raise an issue of researcher bias (Creswell & Creswell, 2020; Merriam & Tisdell, 2016). The detailed examination of the dataset may introduce unconscious bias between the human and machine-generated latent topics. Familiarizing myself with the dataset in detail, as required for the content analysis component, may allow me to predict topics emerging from the machine-generated semantic topic lists. Consequently, I engaged in the following actions to guard against this potential issue:

- I performed the machine learning component first. This allowed me to view the data through a decontextualized lens, which had fewer terms I may recognize and be influenced by, and preserved some novelty in the data because all I saw were higher-level summaries generated by the computer.

- I asked a second human reader (my dissertation advisor) to examine the outputs as a sanity check. My second reader examined the machine learning-generated semantic topics in phase 2 and assigned 2-3 labels to each topic. I then compared those labels with the ones I generated and saw how much overlap exists between topic labels. I also debriefed with the alternate reader to discuss whether they thought their topic labels were congruent with mine. Some labels were adjusted to accommodate our shared understanding of the topic.

While eliminating all bias is unrealistic, these precautions are a reasonable response to take.

After collecting and analyzing the data for each research question, I reflected and wrote a review in my field notes, looking at the entire body of work as a collective. It was important to reflect often, but especially once I finished each research question because these are the points where I was best positioned to identify changes in my thinking between the start and the end of each question. My conclusions were the latent themes I selected for each concept and my observations and conclusions.

Phase 6 - Presenting Results

Presenting Results (phase 6) is where all my analysis came together. Once I decided upon the latent themes, I created tables for qualitative analysis. These methods provide more data and help construct a richer picture of the research findings. Table 6 compared the topics developed by manual and LDA topic extraction methods using the pilot ISER data. These tables provided an overview of the findings under the first two research questions and illuminated how well the themes aligned between the two methods. I used these tables to compare whether there were overlaps in the Manual and Machine Learning columns for the same theme, for example.

Table 6

Latent Topics by Method, Pilot ISER Data

Latent Theme	Theme Present Under Manual Method	Theme Present Under Machine Learning Method
Mission Preeminence	"Mission Preeminence"	"Mission Preeminence"
Serving Diverse Sets of Students	"Serving Diverse Sets of Students"	
Process Driven Education	"Education process drives course and program outcomes"	"Rational Degree Delivery Process"
Traditional College Education Process		"Traditional College Education Process"

My field notes and Tables 4, 5, and 6 were studied to answer research questions 3 and 4:

3. What are the similarities between the topics extracted by automated and manual topic extraction methods?
4. What are the differences between the topics extracted by automated and manual topic extraction methods?

CHAPTER FOUR: RESULTS

The ultimate objective of this study was to explore methods of generating information about colleges' expressions of organizational identity. I believe this study was successful in that aim. Furthermore, unanticipated findings emerged, which may be helpful in other ways. Latent themes did emerge from both automated and manual text extraction methods, and while there was some overlap, there were considerable differences as well. The ISER was a valuable data source, suggesting the utility of using ISERs as data sources going forward. My results suggest that using automated and manual text extraction methods together may yield synergistic—but not interchangeable—results. I will now discuss the key findings for the four research questions.

Key Findings for Research Question 1

Latent themes emerged from the automated text extraction method using the ISER. Recall Research Question 1 - "What are the main identity themes of CA community colleges as revealed by automated topic extraction analysis of ISER Standard IA?" this question focused on identifying latent themes generated by the computer LDA algorithm. In this study, I started with $K = 10$, ran the LDA algorithm, and studied the output with LDAvis (Sievert & Shirley, 2014). This resulted in topic decoherence; topic decoherence was resolved when I reduced the number of topics to 5. The display for this set of topics, as well as a set of key terms for Topic 1, is provided in Figure 13.

Table 7 provides a summary of the five orthogonal latent themes: "Student-Centered Educational Process," "Deliberate Degree Delivery Process," "Stakeholder Coordination of Strategic Planning," "Strategic Planning for Institutional Goals," and "Institutional Planning and Assessment." Below, I briefly discuss each latent theme.

Figure 13

LDAvis for Student-Centered Educational Process

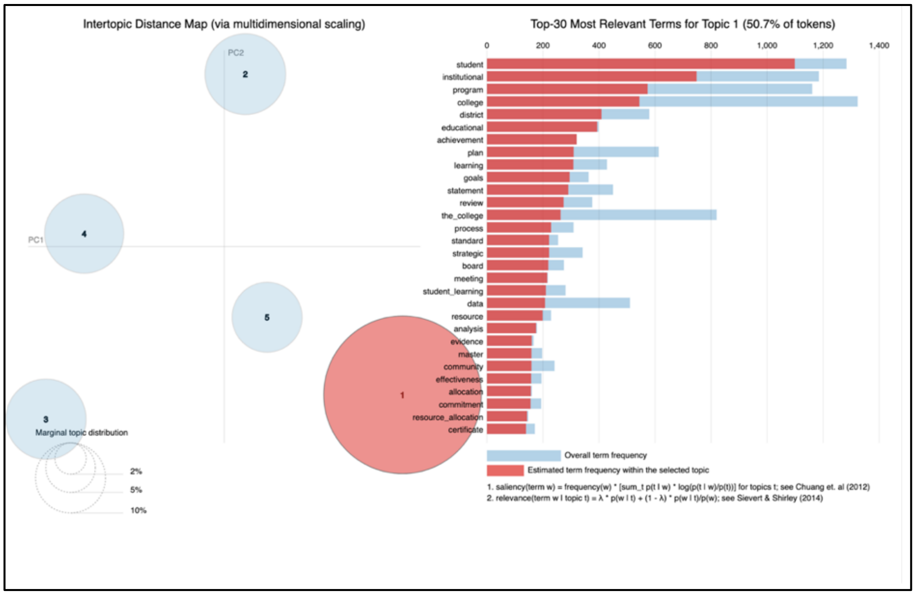


Table 7

Latent Themes Extracted via LDA

LDA Topic No.	Theme Label	Distinctive LDA Terms
1	Student-Centered Educational Process	student, institutional, program, college, district, educational, achievement, learning, goals, review, process, standard, board, meeting, student_learning, analysis, evidence, allocation
2	Deliberate Degree Delivery Process	college, program, student, counsel, learning, data, emp (educational master plan), the_college, college_council, report, minutes, development, faculty, developed, baccalaureate, handbook, activities, model, agenda
3	Stakeholder Coordination of Strategic Planning	institutional, college, data, program, plan, statement, campus, pathways, metrics, trustees, planning_committee, senate, requests, strategic_planning, office, budget, initiatives, guided, board_trustees, academic_senate, facilities
4	Strategic Planning for Institutional Goals	the_college, college, program, institutional, district, course, statements, degree, curriculum
5	Institutional Planning and Assessment	program, institutional, district, process, college, faculty, prioritization, cpc (college planning council), apo (assessment of planning outcomes committee), integrated_planning, manual, identified

LDA Latent Topic 1: Student-Centered Educational Process

The label "Student-Centered Educational Process" was applied to latent topic 1. The terms and a graphic display of their associated amount of lift can be found in Figure 19 in Appendix C. The numerical posterior probabilities and distinctive selection criteria can be found in Table 8. Terms under this theme refer to the college putting students front and center in determining their educational experience.

Table 8

Term Selection Methods for Student-Centered Educational Process

Term	Posterior Probability	Lift	Term Prevalence
student	0.0325		√
institutional	0.0222		√
program	0.017		√
college	0.0161		√
district	0.0121		√
educational	0.0116	√	
achievement	0.0095	√	
plan	0.0092		
learning	0.0091		√
goals	0.0087		√
statement	0.0086		
review	0.0081		√
the_college	0.0078		
process	0.0068		√
standard	0.0066		√
strategic	0.0066		
board	0.0065		√
meeting	0.0064	√	
student_learning	0.0062		√
data	0.0061		
resource	0.0059		√
analysis	0.0052	√	
evidence	0.0047	√	
community	0.0047		
master	0.0047		
effectiveness	0.0047		
allocation	0.0046	√	
commitment	0.0046		

Table 8 summarizes the posterior probability, lift, and term prevalence metrics for Topic 1, as described in the methods section. Terms with higher posterior probabilities (above 0.01) are those terms the LDA algorithm felt the most confident about and are indicated with a white background in the posterior probability column in Table 8, and the grey shading indicated terms with a posterior probability lower than 0.01. All terms with posterior probabilities greater than 0.01 were automatically selected for consideration in topic labeling.

The second column of Table 8 denotes which terms were considered for inclusion in topic labeling based on a high degree of lift. Here, I created an indicator (denoted by a check) of high lift value if there was no blue bar visible – thus indicating this term was important to this topic and this topic only. The resulting solid red bar was helpful in indicating a term capable of contributing a significant degree of distinctiveness to identifying and labeling the topic.

The rightmost column in Table 8 denotes if a term had a high degree of term prevalence. Recall from chapter three that it is similar to lift, except instead of a term existing in only one topic; term prevalence indicates a term predominates (50% or more of that term) within the indicated topic. A high lift term is also a high prevalence term, but only one column is checked to avoid confusion. A high prevalence term may or may not be a high lift term.

Terms that meet the specified thresholds for at least one metric of "aboutness" are bolded. The high posterior probability and high term prevalence terms "student," "institutional," "program," "college," and "district" were selected. The term "educational" was selected because it has high posterior probability and high lift. Additional terms such as "achievement," "meeting," "analysis," "evidence," and "allocation" were selected because of their high lift. The terms "learning," "goals," "review," "process," "standard," "board," "student_learning," and "resource" were selected because of their high term prevalence.

The term "student" supports the student component of the student-centered educational process label. The terms "institutional," "program," "college," and "district" speak to process, while "educational" addresses the type of process. Other high-lift terms—such as "achievement" and "evidence," speak to the educational process, while terms such as "meeting," "analysis," and "allocation" again speak to the process portion of the theme. Finally, the terms "learning," "goals," "review," "standard," and "student_learning" refer to the education process component of this theme label, and "process," "board," and "resource" refer to the process component. Taken together, the terms suggest that a) colleges center students and learning as a primary goal of the institution and b) they marshal institutional resources to deliver this goal.

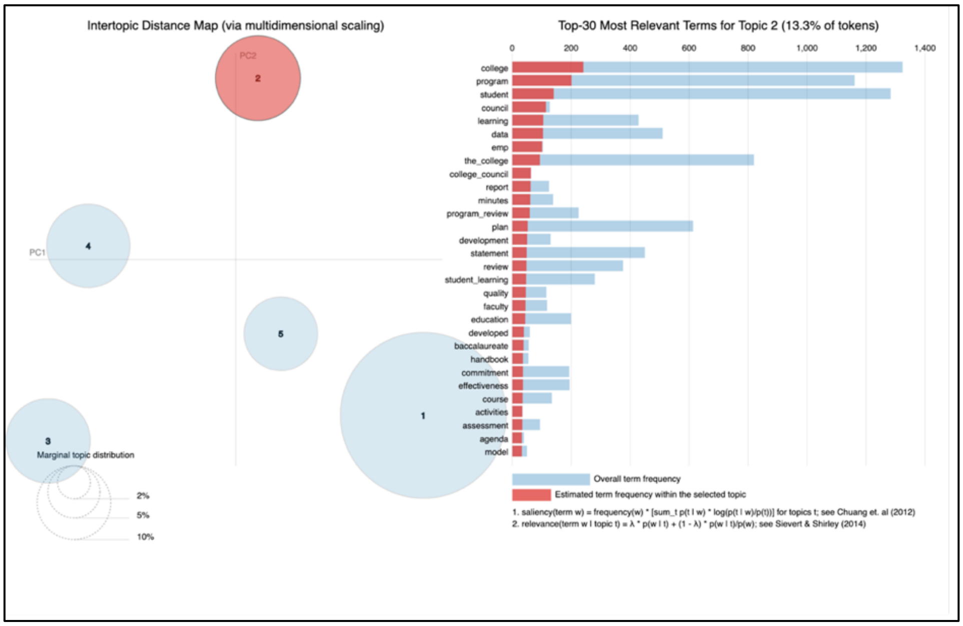
LDA Latent Topic 2: Deliberate Degree Delivery Process

Topic 2 revolves around the “Deliberate Degree Delivery Process.” The terms and their associated amount of lift can be found in Figure 14. The numerical posterior probabilities and selection criteria can be found in Table 9. Terms under this theme refer to an intentional, collaborative, reasoned process for delivering a degree. Community college degrees do not happen; they result from an organized, deliberate, and intentional process.

The terms “college,” “program,” “student,” “council,” “learning,” “data,” “emp,” and “the_college” were all selected in labeling Topic 2 because of their high posterior probability. Two terms are especially noteworthy, “emp” (short for “Educational Master Plan”) because it also possesses high lift and “counsel” because it also possesses high term prevalence. The following terms were selected because they exhibited high lift, “college_counsel,” “activities,” and “agenda.” Finally, the following terms exhibited exclusively high term prevalence, “report,” “minutes,” “development,” “faculty,” “developed,” “baccalaureate,” “handbook,” and “model.”

Figure 14

LDAvis for Deliberate Degree Delivery Process



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Table 9

Term Selection Methods for Deliberate Degree Delivery Process

Term	Posterior Probability	Lift	Term Prevalence
college	0.0272		
program	0.0227		
student	0.0159		
council	0.0128		√
learning	0.0119		
data	0.0118		
emp	0.0114	√	
the_college	0.0406		
college_council	0.0071	√	
report	0.0070		√
minutes	0.0069		√
program_review	0.0067		
plan	0.0059		
development	0.0057		√
statement	0.0055		
review	0.0054		
student_learning	0.0053		
quality	0.0052		
faculty	0.0051		√
education	0.0050		
developed	0.0044		√
baccalaureate	0.0043		√
effectiveness	0.0041		
commitment	0.0041		
handbook	0.0041		√
course	0.0040		
assessment	0.0039		
activities	0.0039	√	
model	0.0037		√
agenda	0.0037	√	

The high posterior probability and high lift term “emp” makes a significant contribution to ‘aboutness’ by speaking towards a formal deliberate degree delivery process. An educational master plan, by definition, is a plan reflecting a formal, deliberate degree delivery process. The high posterior probability, high term prevalence word “counsel” reinforces this idea by referring to a deliberative body (usually faculty-driven) engaged in some educational process activity.

The additional high-lift terms, “college_counsel,” “activities,” and “agenda” reinforced some educational processes, as did the high-term prevalence terms “report,” “minutes,” “development,” “developed,” “handbook,” and “model.” The high-term prevalence word “faculty” speaks to a faculty-driven process. The high posterior probability terms “college,” "program," "student," and "the_college" all refer to some institutional-level educational process (which degrees are). At the same time, including "data" speaks to processes that help inform how well colleges are carrying out initiatives like their educational master plans. Finally, the high posterior probability term "learning" speaks directly to an educational process. The terms "report," "minutes," and "handbook" also suggest artifacts of institutional collaborative processes. The terms "development" and "developed" describe a transformative process – which degrees are. Finally, the high-lift terms "activities," "model," "and "agenda" describe the same generalized institutional process.

The term "baccalaureate" deserves special attention because of its high term prevalence and because it is the only degree type that appears in any of the LDA term lists. Interestingly, the bread and butter degree program for community colleges - the associate degree - does not appear. That is possible because associate degrees are so common that whenever someone mentions a community college degree, the college employees all assume it is an associate degree. A baccalaureate degree, in contrast, is very rare at the community college level, so this term contributes a significant amount of 'aboutness' to this topic. These terms all address the idea of a deliberate process of granting various types of degrees.

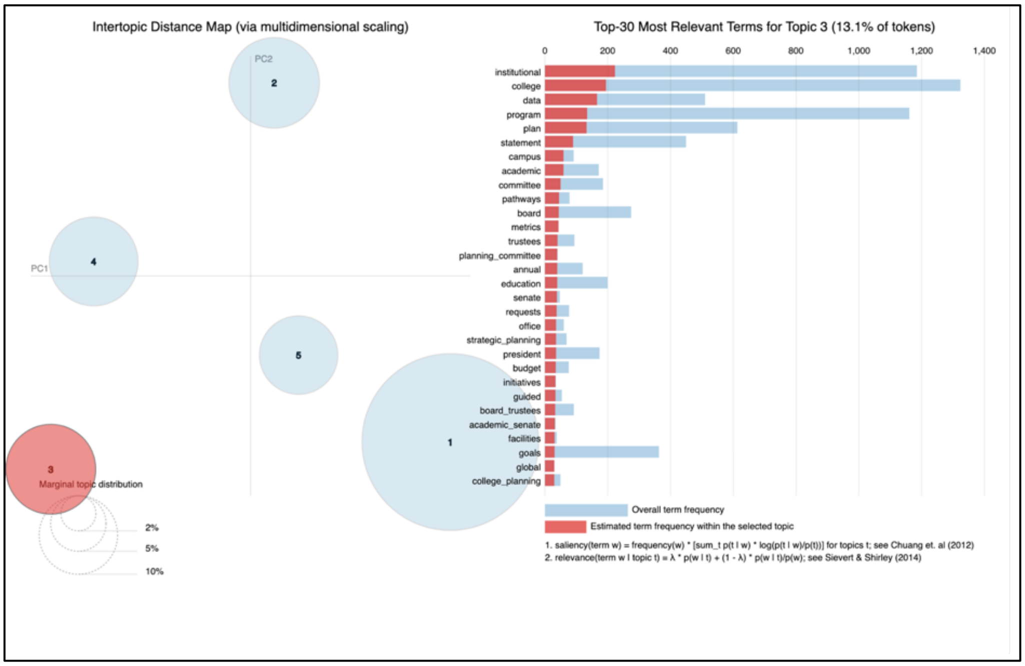
LDA Latent Topic 3: Stakeholder Coordination of Strategic Planning

The label for latent topic 3 is "Stakeholder Coordination of Strategic Planning." The top 30 terms and their associated amount of lift can be found in Figure 15. The numerical posterior probabilities and selection criteria can be found in Table 10. Terms under this theme refer to stakeholder committees involved in ensuring the strategic plan is implemented.

As can be seen in Table 10, there are several high posterior probability terms in this topic, "institutional," "college," "data," "program," "plan," and "statement." There were no high posterior probability terms with either high lift or high term prevalence. The following terms exhibited high lift only, "metrics," "planning_committee," "initiatives," "academic_senate," and "facilities." The following terms exhibited high term prevalence, "campus," "pathways," “trustees,” “senate,” “requests,” “strategic_planning,” “office,” “budget,” “guided,” and “board_trustees.”

Figure 15

Latent Theme 3 - Stakeholder Coordination of Strategic Planning



The high posterior probability terms “institutional,” “college,” “data,” “program,” “plan,” and “statement” all speak to some institutionalized process, as do the high lift terms “metrics” and “initiatives.” Much of the distinctiveness of this topic came from some high-lift terms: “planning_committee,” “facilities,” and “academic_senate” as well as some high-term prevalence terms: “trustees,” “senate,” and “board_trustees.” The high-term prevalence terms “guided” and “pathways” are unusual because they can refer to either a committee or a program (both definitions fit here). The strategic plan aspect is derived from “plan,” “planning_committee,” “initiatives,” “academic_senate,” “senate,” “strategic_planning,” “budget,” and “board_trustees.” A process is implied with the terms “data,” “metrics,” “initiatives,” “requests,” “office,” and “budget.”

Table 10

Term Selection Methods for Stakeholder Coordination of Strategic Planning

Term	Posterior Probability	Lift	Term Prevalence
institutional	0.0255		
college	0.0223		
data	0.019		
program	0.0154		
plan	0.0152		
statement	0.0103		
academic	0.0068		
campus	0.0068		√
committee	0.0057		
pathways	0.0051		√
board	0.005		
metrics	0.0049	√	
trustees	0.0046		√
education	0.0045		
annual	0.0045		
planning_committee	0.0045	√	
senate	0.0044		√
requests	0.0043		√
president	0.004		
strategic_planning	0.004		√
office	0.004		√
budget	0.0039		√
initiatives	0.0038	√	
guided	0.0038		√
board_trustees	0.0037		√
academic_senate	0.0036	√	
goals	0.0035		
facilities	0.0035	√	
strategic	0.0034		
support	0.0034		

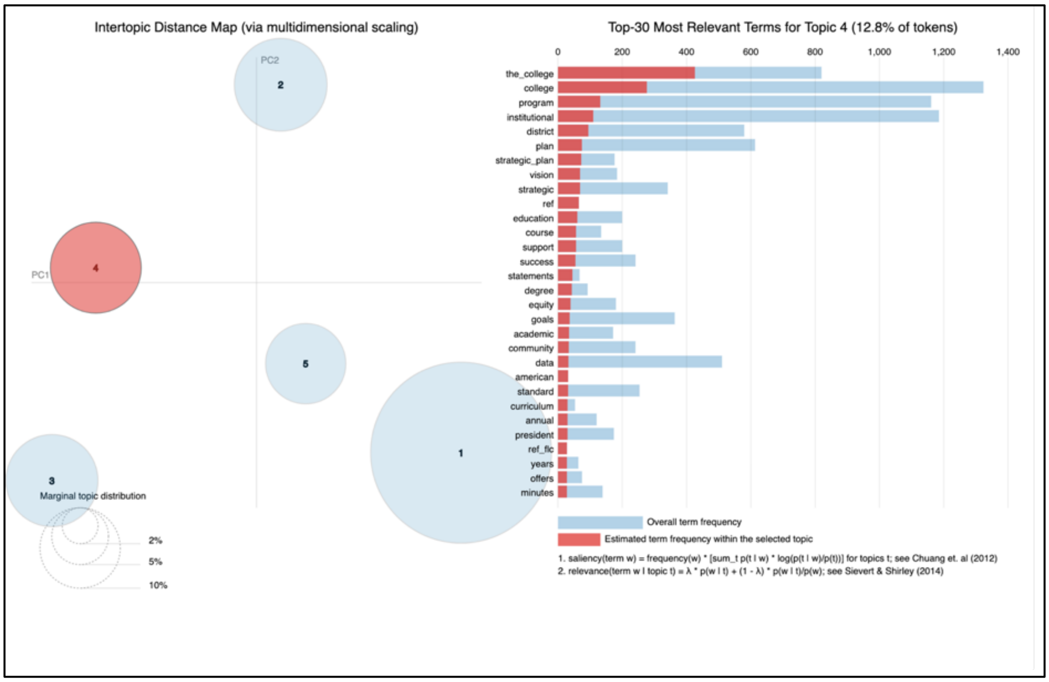
This topic is unusual because there is no one word strongly suggests a label. Instead, groups of words taken together give it cohesiveness. Notice, for example, the large number of stakeholder terms: “planning_committee,” “academic_senate,” “trustees,” “senate,” “strategic_planning,” “board_trustees,” “and academic_senate.” The term “facilities” could refer to a stakeholder group depending on how it is used within the ISER. Also note that another stakeholder, “president,” appears in the term list for this topic but does not meet any of the selection criteria. Taken together, these terms enumerate the various stakeholders for a community college and that, coupled with the various planning terms situate this topic squarely in the realm of term selection methods for stakeholder coordination of strategic planning.

LDA Latent Topic 4: Strategic Planning for Institutional Goals

"Strategic Planning for Institutional Goals," the label for Theme 4, addresses the college's vision of what it wants to do. This topic suggests a long-term focus and a process for getting there. This more aspirational theme answers the question, "How are we going to get to where we want to go from here?" Figure 16 and Table 11 are the significant graphics for topic 4.

Figure 16

Latent Theme 4 - Strategic Planning for Institutional Goals



The high posterior probability, high term prevalence word "the_college" suggests a topic with significant institutional components. This is reinforced by the other high posterior probability terms, "college," "program," "institutional," and "district." The high lift terms were not useful; "ref" appears to be a preprocessing artifact, while "american" is likely part of the name for American River College. "Statements" and "curriculum" suggest a process directed toward some goal. Additionally, "course," "degree," "statements," and "curriculum" all relate to the goals of the institution.

This topic did not hold together as well as the other topics in this study, but part of that may relate to the term selection criteria. The first four posterior probabilities in the “low’ probability text were: “plan,” “strategic_plan,” “vision,” and “strategic.” These terms also suggest the theme of this topic – especially because they were all grouped together, yet they did not meet the selection criteria.

Table 11

Term Selection Methods for Strategic Planning for Institutional Goals

Term	Posterior Probability	Lift	Term Prevalence
the_college	0.05		√
college	0.0324		
program	0.0154		
institutional	0.0129		
district	0.0111		
Plan	0.0087		
strategic_plan	0.0085		
Vision	0.0081		
strategic	0.0081		
Ref	0.0076	√	
education	0.0071		
support	0.0066		
course	0.0066		√
success	0.0064		
statements	0.0053		√
degree	0.005		√
Equity	0.0046		
Goals	0.0043		
academic	0.004		
community	0.0039		
Data	0.0038		
standard	0.0037		
american	0.0037	√	
president	0.0035		
curriculum	0.0035		√
annual	0.0035		
Offers	0.0033		
years	0.0033		
minutes	0.0033		

LDA Latent Topic 5: Institutional Planning and Assessment

“Institutional Planning and Assessment,” the label attached to Topic 5, reflects a focus on the planning process and evaluation with a commitment towards improvement. Significant LDA terms include “faculty,” “prioritization,” “cpc (college planning council),” “apo (assessment of planning outcomes committee),” “integrated_planning,” “budget,” and “manual.” This topic is not as strategic and is more tactical-oriented compared to other topics.

The high posterior probability terms “program,” “institutional,” “district,” “process,” and “college” clearly indicate that an institution is the subject of this topic. The high-lift terms “cpc (college planning council),” “apo (assessment of planning outcomes committee),” and “manual”, as well as high-prevalence terms like “prioritization,” “integrated_planning,” and “identified,” clearly relate to institutional planning activities and were significant in determining the theme of this topic. The term high-prevalence term “faculty” also situates this topic in an institutional planning context rather than a strategic planning context because faculty have a significant degree of involvement at that level of planning.

An assessment dimension of this theme is suggested by the terms “manual” and “identified.” It is also interesting to note that the terms “objectives,” “integrated,” “master,” and “master_plan” were all present in the term list returned by the LDA algorithm but did not make the criteria for inclusion.

Figure 17

Latent Theme 5 - Institutional Planning and Assessment

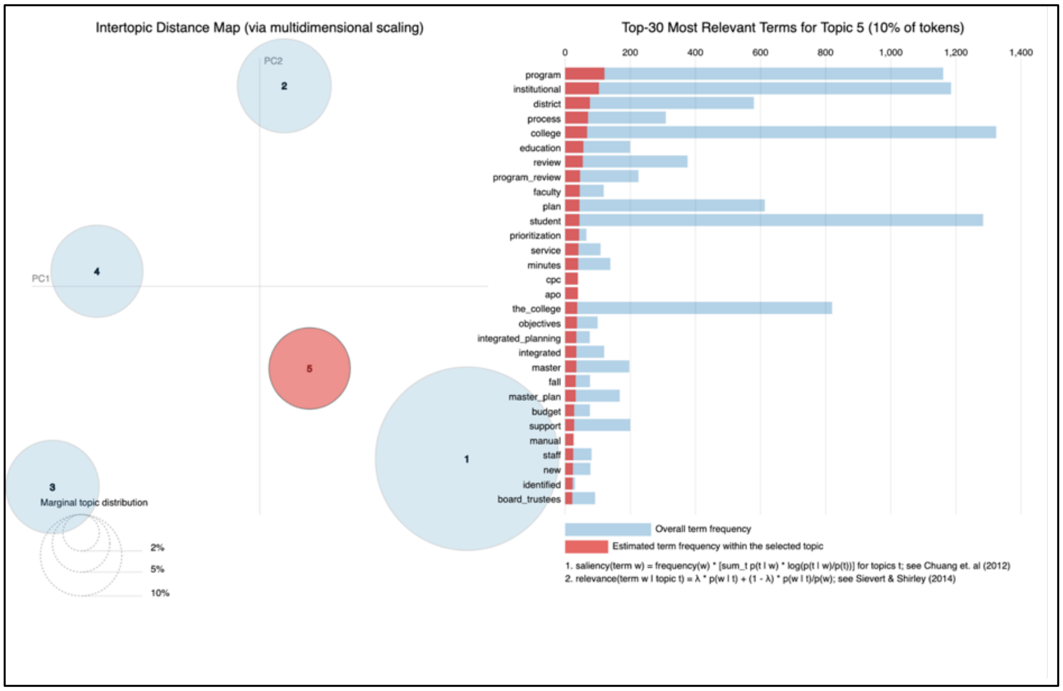


Table 12

Term Selection Methods for Institutional Planning and Assessment

Term	Posterior Probability	Lift	Term Prevalence
program	0.0182		
institutional	0.0156		
district	0.0114		
process	0.0106		
college	0.0102		
education	0.0084		
review	0.0082		
program_review	0.0069		
faculty	0.0067		√
student	0.0066		
plan	0.0066		
prioritization	0.0064		√
service	0.0062		
minutes	0.006		
cpc	0.0059	√	
apo	0.0059	√	
the_college	0.0056		
objectives	0.0054		
integrated	0.0053		
master	0.0053		
integrated_planning	0.0053		√
master_plan	0.0049		
fall	0.0049		
support	0.0042		
budget	0.0042		
manual	0.0039	√	
staff	0.0037		
new	0.0036		
identified	0.0034		√
president	0.0033		

The five LDA latent topics extracted from Community College ISERs revealed process topics revolving around students, educational planning, and assessment. These topics are:

- **Student-Centered Educational Process:** This topic emphasizes the importance of putting students at the center of the educational process and providing them with the resources they need to succeed.
- **Deliberate Degree Delivery Process:** This topic focuses on the intentional and collaborative process of delivering degrees at community colleges.
- **Stakeholder Coordination of Strategic Planning:** This topic highlights the

importance of involving stakeholders in the strategic planning process and ensuring that the plan is implemented effectively.

- **Strategic Planning for Institutional Goals:** This topic addresses the college's vision of what it wants to do and how it plans to achieve its goals.
- **Institutional Planning and Assessment:** This topic reflects a focus on the planning process and evaluation within community colleges.

While it is interesting that these unexpected themes emerged from the LDA topic extraction process, they are also disappointing because they are not organizational identity themes. Recall that one of the tenets of OI is that the identity elements need to be 'central,' 'enduring,' and 'distinctive' (Albert & Whetten, 1985; Gioia & Hamilton, 2016, p. 25; Whetten, 2006). However, not all these topics are 'distinctive'. Topic 1 is present in all the colleges in the sample (not distinctive), while topic 5 is only present in five collegesI; it is unclear that this represents a topic distinctive enough to represent organizational identity. Further details can be found in Appendix B: College Distinctiveness by Posterior Probabilities.

Albert and Whetten (1985; 2006) also suggested that identity-based claims occupy specific social and geographic spaces. While the above themes involve social actors, they do not seem to occupy social spaces. Furthermore, they do not seem to occupy geographic spaces either. Finally, Whetten (06) felt identity themes were more visible when the organization faced "profound fork in the road choices" (p.221). None of the themes listed above seem to be profound choices; in fact, they seem to be the opposite – regular and normal processes an organization would be expected to perform during its day-to-day business. It does not appear that any of these themes meet the criteria (as best we can tell here and now) for possessing attributes of organizational identity. In conclusion, while it is an open question if these latent themes could be considered elements of organizational identity, they nevertheless provide significant (if incomplete) insights into the processes community colleges believe are valuable.

Key Findings for Research Question 2

Research question 2 - “What are the main identity themes of CA community colleges as determined by manual topic extraction analysis of ISER Standard IA?” - focused on using a manual topic extraction method to illuminate latent themes. I used in vivo coding on each ISER, generating an initial set of semantic themes, then re-coded and collapsed those into latent themes.

Table 13 presents key concepts that emerged, along with comments on the central idea and/or the scope of the theme. The column “Latent Theme” refers to the label assigned to the cluster of semantic themes referring to the same concept. The number in parentheses refers to the number of times that concept was identified within the corpus. Concepts are free to appear multiple times within the same college ISER (or not appear at all). I present themes that were present in more than ten instances in the data corpus to focus on the most important elements of identity. The “Central Concept/Scope/Boundaries” attempts to define the latent topic with as little overlap of the other topics as possible. Detailed discussions of the themes begin below.

Table 13

Latent Themes Extracted via Manual/Content Analysis

Latent Theme (#)	Central Concept/Scope/Boundaries
Student Success (41)	The college values the achievements and accomplishments of its students. This is broader than learning. These elements map to a binary state (achieved success/ did not achieve success).
Principles of Social Justice (28)	The college values the ideals of Social Justice.
Common Good (24)	The college encourages the development of skills for the common good and community.
Character Values (15)	Behavior and experiences which foster and honor positive character values.
Innovation Valued (14)	The college encourages the development and expression of new ideas.

Two issues emerged from the sensemaking process with the manual method. The first was the unanticipated discovery of value statements within the response to the standard. I had anticipated several colleges including text from their mission and vision statements, and I was expecting that to become identity data. I did not expect the inclusion of values statements – which enumerated ideas and concepts valued by the college. These immediately became valuable because it was easy to see if the college indicates this is an idea or concept they find valuable, then it may also be an identity element.

The other unanticipated (but in retrospect, perfectly foreseeable) issue was handling the relative frequencies of the manually generated topics. The goal was to look for identity elements, so it seems logical to assume elements of identity that were shared across colleges would appear in the corpus more than some number of times. However, it quickly became apparent that some minimum threshold would be needed to separate the signal from the noise (reoccurring ideas versus ideas that only appeared a few times). While working with the data, it was decided that latent themes possessing more than ten in vivo codes would be considered a sufficiently strong signal for inclusion in the results. This eliminated some potential themes, such as Pedagogy (1), Employment (1), and Student Transformation (8). It was purely a coincidence that the manual and automated methods produced the same number of latent themes.

Manual Theme 1: Student Success

The first theme, Student Success, was the most common component of institutions' identities – appearing 41 times in the ISERs. Colleges articulated goals for student success along several distinct measures. Several discuss success in achievement terms based on degree, transfer, or certificate receipt. For instance, Irvine Valley College notes, "we promote access, success, and equity to meet each student's goals of skills development, certificate, associate degree, transfer, or personal enrichment." (Irvine Valley College, 2017, p. 86) another ISER states, "The mission underscores the college's commitment to broad forms of student success … in terms of completing transfer pathways, degrees, and or certificates" (Butte-Glenn Community College District, 2021, p. 42). While the goals each college mentions are broader than degree receipt, this is clearly a central measure of success for many colleges.

Other colleges focused on the contribution the college could make to the student's future career success. For instance, "College of the Sequoias is … focused on student learning that leads to productive work, lifelong learning, and community involvement" (Sequoias Community College District, 2018, p. 98). Finally, some colleges included open-ended and unspecified student success (in addition to more specific outcomes) as a benefit of attending their institution and invited the student to find their own definition of success, i.e., "CRC's Mission, Vision and Value statements describe the college's broad, educational purpose … to earn certificates or degrees, transfer to other educational institutions, or attain other lifelong, academic or career aspirations" (Consumnes River College, 2021, p. 20).

These quotes, and many others, displayed how broad a concept of student success is to community colleges, encompassing academic achievement, personal development, and social well-being. Notice that the concept is not limited to any specific subject area but is a foundation for lifelong student goals. This is one of the strengths of the California Community College system: a multitude of paths to student-defined success. Furthermore, it is not a uniform, mass-produced physical good but an individualized process.

Interestingly, these quotes reflect a shift in how student success is conceptualized. Historically, student success was often thought of as individual personal achievement – explicitly grades (Straumsheim, 2017). However, the above quotes suggest that community colleges are reconceptualizing their relationship with their students and communities as partners, and they are increasingly committed to providing students with the support they need for success.

Manual Theme 2: Principles of Social Justice

The second theme, Principles of Social Justice, appeared in the ISERs 28 times. Colleges enumerated their Principles of Social Justice goals along three major axes: diversity, equity, and mutual respect. Diversity is conceptualized very broadly with few, if any, qualifiers (such as gender or race), e.g., "The College is dedicated to … fostering diversity … (Cypress College, 2017,

p. 83) or “West [Los Angeles College] fosters a diverse learning community … (West Los Angeles College, 2023, p. 113). The equity axis focuses on ideals of fairness or access, e.g., “Las Positas College provides an … equity-focused environment that offers educational opportunities …” (Las Positas College, 2022, p. 73) or “[Consumnes River College’s] values are demonstrated through [our] commitment to equity and social justice in student outcomes …” (Consumnes River College, 2021, p. 21). Several colleges recognized that social justice would not be possible without mutual respect and included this element in their ISERs; for example, “American River College strives to uphold the dignity and humanity of every student and employee.” (American River College, 2021, p. 52), or “[Palomar College is] guided by [our] core values of mutual respect and trust through transparency, civility, and open communications.” (Palomar College, 2022, p. 53) or “We demonstrate a commitment to the value of each individual through trust, cooperation, and teamwork” (Laney College, 2020, p. 43).

This topic reflects the belief that all students, regardless of their background, deserve a fair and equitable education. Colleges valuing this theme believe a college is where students should feel valued, included, and respected. These colleges understand that recognizing the principles of social justice is essential for creating a fair, inclusive, and welcoming environment. This topic, in particular, would be interesting to examine over time, given the Chancellor’s office equity initiatives such as Vision for Success and the passage of the CA legislatures’ AB705 and 1705. The method employed in this dissertation may be able to determine if those equity efforts had an impact on community college identity.

Manual Theme 3: Common Good

Common Good is the third topic. It appears in the ISERs 24 times. This theme is embodied in a few ways: community membership, economic development, cultural awareness, and environmental justice. Community membership acknowledges that the college is situated within and influences a specific community, “The vision of the college is to empower students and employees to strengthen the cultural, social, economic, and environmental well-being of their communities" (Consumnes River College, 2021, p. 20). Economic development acknowledges the economic impact colleges have within their communities. "The college is dedicated to supporting the success of our students, fostering diversity, enriching society, and contributing to the economic development of our community and beyond" (Cypress College, 2017, p. 83). Notice this is greater than individual financial success for the student; colleges are also discussing their role as engines of economic growth.

Cultural awareness is similar to community membership in that it acknowledges that the college is situated within something larger than itself. However, in the case of cultural awareness, the college is acknowledging it is part of a larger cultural heritage, "The College's unique mission statement honors the culture of each of the three locations - the main campus, Colusa County Campus (CCC), Lake County Campus (LCC) - and addresses varying populations each location serves while providing an overarching mission encompassing all three Woodland Community College sites" (Woodland Community College, 2018, p. 61).

Lastly, environmental justice recognizes the role the college has in fostering environmentally sustainable practices and providing stewardship of the natural environment. For instance, Butte-Glenn Community College district states that "We promote and model practices that will result in positive outcomes for our human and natural environments and the long-term viability of the College" (Butte-Glenn Community College District, 2021, p. 43).

This broad topic stipulates that colleges are a force for good in their communities. The themes above provide a good example of the many ways this can be accomplished: economic development, the wise use of resources, and acknowledging the preexisting culture. The common idea binding these themes is one of community.

Manual Theme 4: Character Values

Our fourth topic is Character Values, and it appeared in 15 colleges. Colleges defined this theme along four dimensions: accountability, excellence, integrity, and leadership. Accountability is the willingness to hold college employees responsible for student success, "we are individually and collectively responsible for achieving the highest levels of performance in helping students acquire the necessary skills and abilities to earn associates degrees, certificates, transfer, and career preparation. We continually evaluate ourselves in an effort to improve our effectiveness and efficacy in meeting the educational needs of our community" (Laney College, 2020, p. 43). Excellence in teaching reflects the institution's willingness to hold itself to the highest standards in instruction, as exemplified by ISERS stating: "Palomar College is … guided by our core values of … excellence in teaching, learning, and service" (Palomar College, 2022, p. 53) and "Butte College … provides quality education and support services that are continuously evaluated and improved …" (Butte-Glenn Community College District, 2021, p. 42). Several colleges spoke to integrity, emphasizing values such as ethics, truth, and honesty: one ISER states that "Education is the reason our institution exists. To this end, we value innovation, professionalism, integrity, and responsible stewardship." (Modesto Junior College, 2017, p. 63) while another affirms that "Integrity as the foundation for all we do" (Palomar College, 2022, p. 53).

This topic reflects the belief that certain inner qualities (such as accountability, excellence, integrity, and leadership) are essential for ethical and moral behavior. It is interesting to note that ethical and moral behavior attributes are not just for students. Hartnell College, for example, applies character values to the college itself, "Ethics and Integrity. We commit to respect, civility, honesty, responsibility, and transparency in all actions and communications" (Hartnell College, 2019, p. 49).

Manual Theme 5: Innovation Valued

Innovation values were the last latent topic in 14 of the colleges sampled. This topic consists of a single dimension: innovation. This represents the idea of bringing change to aspects of the college, "we encourage and support creativity, collaboration, and risk-taking. We foster and promote innovation in the design, development, support, delivery, and management of all programs and services." (Laney College, 2020, p. 43), "We provide a dynamic and innovative learning environment for diverse learners of all ages, backgrounds and abilities." (Irvine Valley College, 2017, p. 86), "We provide a dynamic, innovative undergraduate, educational environment for the ever-changing populations and workforce needs of our community." (Modesto Junior College, 2017, p. 63), and "We are committed to sharing and exploring new ideas through collaboration, respect for diversity, promoting equity, and professional development. (Fresno City College, 2018, p. 123), and "Butte College is a student-centered learning institution, which provides quality education and support services that are continuously evaluated and improved to prepare students to be productive members of a diverse, sustainable, and ever-changing global society" (Butte-Glenn Community College District, 2021, p. 42). Most noticeable about this dimension is its breadth. Innovation ranged from novelty in college programs and services to the learning environment through methods such as collaboration and professional development to the concept of continuous improvement. This single label covered a great deal of conceptual ground.

This topic refers to introducing new ideas to the college – usually in the form of educational practices, programs, or policies to improve student learning. Usually, these innovations are discussed in alignment with the college's mission or goals statement, implying that these innovations are practical, sustainable, and scalable.

The five topics manually extracted from community college ISERs address the most frequent self-reported values the colleges believe are important. These five themes are:

- **Student Success:** These quotes, and many others, displayed how broad a concept of student success is to community colleges, encompassing academic achievement, personal development, and social well-being.
- **Principles of Social Justice:** This topic reflects the belief that all students, regardless of their background, deserve a fair and equitable education.
- **Common Good:** This broad topic stipulates that colleges are a force for good in their communities.

- **Character Values:** This topic reflects the belief that certain inner qualities (such as accountability, excellence, integrity, and leadership) are essential for ethical and moral behavior.
- **Innovation Values:** This topic refers to the idea of introducing new ideas to the college – usually in the form of educational practices, programs, or policies to improve student learning. We know these are elements the colleges value, but we do not know if they are elements of organizational identity.

These themes are closer to the types of topics I expected to uncover in this study. However, I hesitate to label them organizational identity themes, although I believe they are closer to that concept than the LDA themes. The big issue is that these themes come up short if we apply the same criteria we used for the LDA themes. Revisiting the fundamental tenants of OI identity elements as 'central,' 'enduring,' and 'distinctive,' (Albert & Whetten, 1985; Gioia & Hamilton, 2016, p. 25; Whetten, 2006) we are again faced with the same issues in determining if the themes are 'central' or 'enduring' – the dataset is not set up to address these questions. However, a hypothesis testing type of test may be too strict a standard for social research. These themes do seem 'distinctive' because many of them are explicitly mentioned by the individual colleges; however, we still do not know if they are 'distinctive' enough. These themes do not seem to fit any geographic spaces (Albert & Whetten, 1985), but there does seem to be some correspondence with social spaces (Albert & Whetten, 1985). It is not outside the realm of possibility to see a college using elements of 'Principles of Social Justice' to stake claims in a social space, for example.

Additionally, it is not inconceivable to imagine the college invoking elements of 'Student Success, Principles of Social Justice, or Common Good" when faced with "profound fork in the road choices" (Whetten, 2006, p. 221) either. These reasons suggest that the themes here are closer to themes of organizational identity, but some issues still need to be resolved to claim that it is, indeed, what they are. However, despite that, these themes still may provide valuable insights into the ideas specific colleges hold dear. This work was motivated by OI theory, and the two analysis methods used in this study did illuminate aspects of college OI.

One unexpected finding from Table 13 was a topic eliminated because of too little support – pedagogy. The best summary of this theme is a quote by Hartnell College, “We commit to excellence in teaching …” (Hartnell College, 2019, p. 49). The surprise was that this theme was not more popular as education is the primary purpose of community colleges (but, as this study indicates, not the only purpose), so one would have assumed it would have ranked prominently in the sample.

Key Findings for Research Question 3

In Research question 3 - “What are the similarities between the topics extracted by automated and manual topic extraction methods?” I explore the overlap between the latent topics produced by the automated and manual methods.

Table 14 displays all the latent themes and illustrates the method that generated them. The summary tables from the previous two research questions show that automated (Table 7) and manual (Table 13) text extraction methods can extract latent themes from the corpus.

Table 14

Comparison of Latent Themes Generated by the Two Text Extraction Methods

Latent Theme	Manual Theme	LDA Theme
Character Values	Character Values	
Common Good	Common Good	
Innovation Valued	Innovation Valued	
Principles of Social Justice	Principles of Social Justice	
Student Success	Student Success	Student-Centered Educational Process Deliberate Degree Delivery Process
Stakeholder Coordination of Strategic Planning		Stakeholder Coordination of Strategic Planning
Strategic Planning for Institutional Goals		Strategic Planning for Institutional Goals
Institutional Planning and Assessment		Institutional Planning and Assessment

The “Latent Theme” column on the left was generated by comparing the key themes from the automated and manual methods. If the two methods generated similar themes, one would expect to see a smaller table with fewer blank cells. However, in this case, the themes were quite different, and there was relatively little overlap between the various themes. Even where there was some overlap (e.g., “Student Success”), the LDA latent topic was much less inclusive than the manual topic.

Examining Table 14 reveals only one common theme between the two methods. This overarching theme is labeled Student Success, joining the Student Success theme from the manual analysis with the Student-Centered Educational Process and Student Success and Deliberate Degree Delivery Process themes from the automated method. Surprisingly, although automated and manual methods extracted topics from the same texts, the overlap was relatively small.

Comparing Student Success and Student-Centered Educational Process topics revealed that both topics addressed the same subject (students) and the college's delivery of education to them. However, what precisely the education was differed. LDA presented abstractions of educational processes such as “learning,” “goals,” “educational,” and “achievement,” but it did not articulate specifically what these were. Manual text extraction uncovered the specifics such as “lifelong learning,” “employment,” and “critical thinking,” but they were not summarized. The methods highlighted the same concept, but the way they did it differed. The common ground only became visible once the researcher made the mental leap by recognizing that the two methods were discussing the same thing but approaching it from different directions.

Student Success and Deliberate Degree Delivery Process primarily overlap in the benefit they provide students – a degree. LDA extracted “baccalaureate” while the manual methods extracted topics in which degrees were a component, such as “student achievement” (“[At Irvine Valley College], we promote access, success, and equity to meet each student’s goals of skills development, certificate, associate degree, transfer, or personal enrichment" (Irvine Valley College, 2017, p. 86)) or student empowerment" "CRC's Mission, Vision, and Value statements describe the college's broad, educational purpose to empower our diverse students to earn certificates or degrees, transfer to other educational institutions, or attain other lifelong, academic or career aspirations" (Consumnes River College, 2021, p. 20). A similar conceptual leap to what was discussed in the prior paragraph was required by the researcher to recognize that these two themes had overlapping components. Furthermore, like the previous paragraph, both themes discussed the same subject, "students."

We can observe another similarity between the two sets of latent themes if we move outside the topics presented in Table 14 and examine the processes involved. Unsurprisingly, there was a significant amount of researcher labor in producing the latent themes for the manual methods. It was a qualitative process that possessed all the strengths and weaknesses one would expect (Merriam & Tisdell, 2016). The surprise came from the degree of researcher labor involved in making judgment decisions about the automated processes. Even with this quantitative method, the researcher was also called upon to make many judgments (what terms to preprocess, what text to include in the corpus, how many topics to use, and what terms are distinctive for this topic, etc.). In this regard, LDA had more in common with qualitative methods than traditional quantitative methods like correlation calculations or plotting a regression line. Nevertheless, this technique might be useful to unambiguously cluster colleges based on a constellation of identity themes. If that could be done, these constellations could be examined to see if there was any association with student outcomes.

Some conceptual overlap exists between the automated and manual groups of latent topics. Both groups are very much centered on students and refer to mindful and deliberate educational processes. Intentionality on the college's part shines through both groups of latent topics, and they also missed terms that could be elements of OI. This work extends what is known in the field by generating data from the ISERs using two different methods.

Key Findings for Research Question 4

In research question 4 - “What are the differences between the topics extracted by automated and manual topic extraction methods?” I examined the variation between latent topics.

As previewed in the section above, the orientation of the groups of themes between manual and automated methods was quite different. As alluded to above, the distinct LDA latent themes reflected the nature of the ISER (a primarily administrative document for administrative purposes), and the latent topics (e.g., “Stakeholder Coordination of Strategic Planning”, “Institutional Planning and Assessment”) have an academic process management feel to them. The LDA terms are focused on actionable activities and were relatively specific in nature. They revolved around concepts administrators and managers (and accrediting agencies) would find useful.

In contrast, the distinct manually extracted latent themes (e.g., “Common Good,” “Social Justice”) were more directed towards more intangible entities: values and concepts. These terms were broader and more comprehensive than those extracted by automated methods. They were much more generalized in nature and covered a broader conceptual area. The manually extracted topics arguably do a better job of capturing the identity of the institutions per se, while the automatically extracted topics arguably capture the means institutions use to enact those identities.

Who the actor was in each theme is another area of distinction between the two methods. The latent topics generated by the computer all situated the college (or the district or institution) as the actor. The topics referred to getting something done, and the college was doing it. This is likely one reason why every latent topic had “college” or “the_college” scoring very high on the posterior probability. This is completely understandable, given the nature of the ISER. However, the manual latent topics possessed far more eclectic actors. For example, under student success, “As a ‘student-centered, open-access community college,’ all members of the community are welcome to enroll in classes" (American River College, 2021, p. 20). The student enrolls in the class. The college is passively waiting. However, this is different with innovation, "Butte College is a student-centered learning institution, which provides quality education and support services that are continuously evaluated and improved to prepare students to be productive members of a diverse, sustainable, and ever-changing global society" (Butte-Glenn Community College District, 2021, p. 42). In this text, the benefit accrues to the student, but the college is the entity with agency because the college must perform innovative acts. Yet, the agency of the common good theme is different from either of these, "The College's unique mission statement honors the culture of each of the three locations - the main campus, Colusa County Campus (CCC), Lake County Campus (LCC) - and addresses varying populations each location serves while providing an overarching mission encompassing all three Woodland Community College sites." (Woodland Community College, 2018, p. 61), here agency flows from the local culture to both the college and the students.

I did not expect the differences between the two methods to be as significant as they were; I expected the topics extracted to be quite similar. However, there were a number of differences between the two methods, which likely explain this result. The first difference involves the focus that the various methods employed. Even though the same data set was used, the methods did not treat the dataset the same. The automated method examined all the text under standard I.A: Mission and constructed groups of co-occurring terms accordingly. The manual method was focused by the unexpected appearance of values statements in standard I.A.1. In a project focused on organizational identity, values statements should be an unexpected and welcome boon. However, the unanticipated consequence of this is that most of the researcher's attention was focused exclusively on standard I.A.1: Description of Educational Purpose, rather than on all the standards under section I.A. This was not intentional; it simply became very clear early on that there was very little relevant material for OI in standards I.A.2: Data Driven Institutional Evaluation, I.A.3: Program and Service Mission Alignment, and I.A.4: Published Mission

Statement. This represented an advantage for manual coding methods: A human researcher has the research question to guide them when reviewing the data, and in this case, the research question limited the human to searching for terms embodying OI. The researcher could keep this 'in mind' when working and only examine potential candidate terms. While some machine learning methods might be trained to filter data like a human researcher, LDA is not one of those such methods. The computer algorithm ingested all the material it was given and searched all possible term combinations for any theme lurking there. There was no way to limit the LDA algorithm to look for elements of identity except through the corpus itself. LDA did not have the ability to keep OI 'in mind' when it was extracting topics. It found themes that arose because the same terms co-occurred. Consequently, the two methods possessed two different methods for filtering the dataset.

A second difference involved the way relationships between terms constructed meaning. The LDA algorithm needed physical distance between explicit terms to illuminate their relationship. The human researcher can use implied relationships between inferred terms to extract the latent topic.

Analyzing the data for research question four suggests meaningful differences exist between the two sets of latent themes. The latent topics themselves differed, as did the actors within those topics. This makes sense as the two methods differed in how they focused on individual terms, filtered the data, and handled relationships between the terms. This work extends the academic conversation in the field by articulating specific differences between automated and manually produced latent topics.

The findings of this study also have implications for college change agents. The results suggest automated and manual text extraction methods can extract themes from a corpus of ISERs. The two sets of themes will likely share some similarities but will not perfectly overlap. This may be beneficial because researchers using both methods can illuminate more conceptual space of a phenomenon under study than either method alone. The automated text extraction method identified process-based themes using the ISER and will likely prove useful if the researcher is interested in that phenomenon. I observed the manual method was better able to extract themes comprising concepts colleges value. This may serve the original purpose of this study in facilitating organizational change interventions as well as identity themes were expected to do. However, it is hard to say if either method extracted identity themes. Overall, this study's results significantly contribute to a number of academic conversations relating to educational leadership.

CHAPTER 5: CONCLUSION

Several conclusions can be drawn from the results of this study. First, LDA and manual topic extraction can both be used to extract topics from Community College ISERs. This finding is consistent with the literature. Secondly, automated and manual text extraction methods generated different perspectives on the data, but with some overlap. LDA can lead to less predictable outcomes because bias towards a specific phenomenon can only come from the dataset. Researchers cannot simply throw data at the algorithm and expect it to yield the desired results. Elements of college identity might be more visible if a more restrictive document set were used (like mission or value statements, for instance). Finally, this project succeeded in its goal of generating new information for college change agents to use in planning interventions.

Distinctions Between the Latent Topics

Analyzing the data revealed several distinctions between the two groups of latent topics. The LDA topics focused on specific processes rather than outcome values and concepts, with an institutional rather than a more generalized and comprehensive perspective. This makes sense as the ISER is a tool the institution produces to discuss how its processes align with the accreditation standards. The differences in themes uncovered by the automated and manual methods suggest that automated methods are not substitutes for manual methods. An analogy may be one of adding an additional colleague to a qualitative project. This new researcher will have different experiences and, consequently, generate slightly different results.

The literature on the degree of expected common latent theme overlap between automated and manual methods is unclear. Previous studies suggested that the two methods would overlap to a significant degree (De Graaf & van der Vossen, 2013; Muller et al., 2022), but they were silent as to how significant the differences between the topics of the two approaches were. For example, De Graff & van der Vossen (2013) were interested in extracting identity frames from press releases and hypothesized that the better the identity matches with the ‘story’ of the press release, the faster other news media would adopt the press release. Comparing automated and human methods was a secondary concern. The implication was the two methods they used would generate comparable results. In the work by Muller & Associates (2022), studies were clustered using an automated unsupervised learning method and a manual method to determine if the automated clustering process for creating a literature review could replace a manual review creation process. Note that while the LDA approach used in this study is also an unsupervised learning algorithm, neither study explicitly used LDA. Also, note that press releases or published studies are not semi-structured or structured documents. Researchers should not expect LDA and manual analysis methods to generate the same latent topics, even when presented with the same data set. This may not necessarily be bad. This work adds to the academic conversation on this topic by exploring some possible reasons why there may be inherent differences between the two methods.

Another reason the topics differed may be the way the two methods handle the data itself. Some type of filter is necessary to make sense of the data; otherwise, every term will be considered important. For example, the research question focuses human attention on the phenomenon of interest, thereby serving as a filter to restrict unwanted noise. In this study, the concept of OI filtered the researchers' attention on topics that could plausibly be considered elements of organizational identity. LDA lacks this ability to restrict the universe of potential topics and could easily find topics outside of the researchers' interests. This was an impediment in this study but may not be in other studies. The LDA extracted process-oriented latent themes, such as “Student-Centered Educational Process” or “Institutional Planning and Assessment,” which were not concepts the researcher was initially seeking information on, for instance. The semi-structured nature of the ISER provides a way for the researcher to focus the attention of the algorithm on the topics of interest in a manner like the lens provided by the research question to humans. This unexpected finding was not discussed in the higher education literature.

Comparing the various key findings suggests that automated and manual methods of topic extraction complement one another because they appear to focus on different aspects of the corpus. The literature suggested a hybrid method combining manual and automated text extraction could be valuable. This research extends that conversation by suggesting that a fusion of the two methods is not the only way to gain; concurrently using automated and manual methods on the same dataset could also yield rich results.

The ISER is a Useful Data Source

The ISER was a useful data source. The data shows that it is possible to extract important latent themes from college ISERs using automated or manual methods. The document's semi-structured nature made it possible to focus the algorithm on the section of the ISER that could be expected to yield the best results for themes of interest. The literature (Sun et al., 2019) does speak about the utility of applying topic extraction to accreditation reports. Furthermore, a fundamental tenant of text processing is textual data usually requires some form of preprocessing (Grimmer & Stewart, 2013). However, the literature was silent on shaping the LDA model's output using the ISERs' ability to focus the computer on specific areas of interest. This is an unexpected and unintended finding of this study.

For humans conducting manual coding, the semi-structured nature of the document allowed the researchers to zero in on the topics they were looking for. Standard I.A was comprised of four substandards. However, standard I.A.1 was the predominant source of data around college identity, and it contained valuable value statements. The other three sections of the standard contributed very little to the manual analysis because they did not discuss topics directly relating to OI.

Keeping organizational identity in mind was important for generating useful data for the manual extraction process. In fact, examining the data with any kind of lens may be important for a human researcher. If the researcher were just fishing, then in vivo coding would be unbounded. Every noun and verb would need to be coded because the researcher would not know which word would be important. Selecting OI terms was essential to limiting the set of concepts to search for and for guiding the results in a path compatible with the objectives of this study. This finding is consistent with the existing literature. Furthermore, this work extends what is known in the field by generating data from the ISER.

Interestingly, the semi-structured nature of the ISER was good for machine learning algorithms as well. In fact, this may be one reason why the LDA extracted topics were different - there was no way to limit the LDA algorithm to look for elements of identity except through the corpus. LDA did not have the ability to keep OI ‘in mind’ when it was extracting topics. It found themes that arose because the same terms appeared together over and over, suggesting there were latent concepts that were important enough for the college to include in its ISER. These are the broad processes involved in running the college. The latent topics comprise key components of college administration (probably not an exhaustive list) that revolve around executing the college’s mission. The mission informs the strategic direction of the college. The college needs to create strategic plans to achieve its mission. Higher education programs are the colleges’ value statements and must be aligned with the goals and objectives of the college. The college needs to manage its resources purposefully to achieve its strategic goals. Finally, institutional planning and assessment close the loop in the administrative process.

This study has made several contributions to the literature on automated and manual text extraction methods. First, it has shown that both methods can extract latent topics from Community College ISERs. Second, it cautions researchers to be aware that the two approaches will likely generate two different perspectives of the data. Finally, it confirms the literature findings on the utility of the ISER as a data source.

Limitations

The mixed method design of this study is well suited to address the research questions; however, there are also several limitations to consider when interpreting the results. OI was incredibly useful as an organizing and motivating framework. However, the theoretical limitations of the OI framework mean that it is difficult to determine whether the latent themes identified in the study are truly representative of a college's organizational identity. Second, the LDA algorithm is a statistical method sensitive to the values used in various parameters. This means that the results of the LDA analysis may be sensitive to the specific settings used in this study. Third, the potential biases of the ISER as a data source suggest that it is possible that the latent themes identified in the study are not representative of the entire college community. Finally, the impact of COVID-19 suggests a confound may exist in the data between the pre-COVID and post-COVID eras.

Limits Of Organizational Identity

The concept of Organizational Identity has theoretical limitations as a concept as well as operationalization issues. Researchers should be aware of these potential issues.

OI as a theoretical concept may give too much weight to the personification of identity and imply that OI is much more fixed and rigid than it is. Some researchers wonder if there is even a phenomenon called OI. Others believe it is best described as a metaphor and recognize that even this description carries ontological baggage (Foreman & Whetten, 2016, p. 130). This may lend the appearance of undeserved substance. Researchers may think they are investigating rocks, only to discover that they are investigating smoke. Therefore, I recognize the need to exercise caution in claiming to have uncovered organizational identity themes within the California Community College system.

In the ISERs used in this study, a number of colleges published value statements in their ISER. These colleges explicitly said, “We value these ideas ….”. Yet even with that clear and definitive statement, one would be hard-pressed to claim that the ideas the colleges enumerated rose to the level of identity themes. We could firmly claim that the elements identified represented college values since colleges explicitly declared them as such. Perhaps these concepts rise to the level of beliefs? However, it is hard to determine if they are elements of identity. One reason for this may be because OI themes are composed of organizational elements believed to be central, enduring, and distinctive (Albert & Whetten, 1985; Gioia &

Hamilton, 2016, p. 25; Whetten, 2006). One can strongly claim that the themes generated by automated and manual methods in this study are central and distinctive. But there is no way to determine how enduring these themes are from one ISER alone. If the same themes emerged from the college's ISERs created over several accreditation cycles, one could begin to claim these are identity elements.

LDA Validation

Contaminating the results of the LDA method with the in-depth familiarization of the dataset required by content analysis was an initial concern in this study. In the methods section, I described the steps I took to validate the latent themes I extracted by the machine learning approach – primarily to guard against the risk of knowing about the themes from my earlier qualitative work contaminating my analysis of the machine learning results. Researchers must be confident in a method's results, especially when surrendering control to machine learning algorithms. In this study, I worked to increase confidence in my analysis by asking a second reader—my dissertation chair—to review the LDA output and develop labels for the themes; our initial thematic labels were broadly similar, and we conferred to land on the final labels. Moreover, my initial concern was that the latent themes produced by the automated and manual methods would be similar; as this was not the case, contaminating the automated method with prior information gleaned from the manual methods did not seem to be an issue.

However, the differences in themes raised other validation issues instead. Implicit in this study was the idea that the two methods could cross-check and validate each other. However, the differing results in this study suggest that this may not be possible as the outcomes from automated and machine learning methods may be too different. While a potential source of validation may be lost, a new lens to view the data was gained.

Care should be taken to ensure researchers actually possess a new lens. The way the data is preprocessed for the LDA algorithm may introduce confounds. There is no processing of figures, tables, or images; the algorithm only examines text. However, the human researchers saw any figures, tables, or images present in the data. While not focusing on these items specifically, their mere presence could explain why the results differed between the two methods.

Moreover, the LDA algorithm analyzed other data the human researchers did not – bigrams. These are terms comprised of two adjacent words with an underscore connecting them. Creating a manufactured term comprised of two already significant terms may confound results by skewing the theme of a topic. For example, the bigram "academic_senate" appears in topic 3, as do the terms "academic" and "senate." The term "academic_senate" overlaps the component terms as well and may skew topic 3 in an academic governance direction. It is also possible the bigram had little effect because its constituent terms moved the topic in that direction anyway.

A researcher can also inadvertently create a bigram as well. Earlier, I indicated I needed to create a term to indicate the college name, which was uniform across all the ISERs. I hypothesized the college name would provide information about a topic by its placement in the text. However, the term "the_college" was created using the same notation as the creation of bigrams, namely the underscore character connecting the two words. The term "college" is a frequently occurring term, but it may or may not have the same meaning as the name of the college. Having two closely related terms with significantly different meanings may have introduced a confound in the results. I still believe changing the college name to something generalizable throughout the corpus is necessary. However, I suggest future studies replace the term "the_college" with a completely nonsense term that does not appear elsewhere in the corpus.

Frequently occurring terms may also create a confound as well. Terms like "college" and "student" are expected in an ISER and may not provide any information to differentiate topics. Terms occurring less than three times were filtered out of the corpus, but not popular terms. Yet researchers may want to think twice before filtering out popular terms. These terms are common because they are often objects of specific processes. The terms "academic," "senate," and "college" may refer to a very different topic than the terms "academic," "senate," and "student." More research is needed to tease out the effect of very common terms on LDA topics.

A related validation issue is the ease of generating output using computerized tools. LDA is a stochastic algorithm. It uses a random number generator in its calculations. Unless one fixes the random number seed, one will get different results (topics) every time it is run. At first blush, this may seem like a problem, but it is not much different from changing the human researchers for every look at the data in qualitative content analysis. One would not expect two humans to look at complex conceptual data and extract the very same topics (overlap, sure, identical, no) every time.

The problem is that one can keep running the LDA algorithm until one arrives at "desired" results; running the program 1000 times isn't a problem. Lining up 1,000 researchers to perform the same task with content analysis is not feasible. This becomes an even more significant issue when the results of the LDA algorithm are presented as either the 'best' or 'an average' set of latent themes. With any given run, there is no way of knowing how representative the distribution of terms this particular set of terms is. Researchers using manual content analysis would never think to make the claim that "my analysis is the absolute best answer; there can be no other interpretation." However, the quantitative nature of machine learning algorithms makes it very easy to assume that the computer has arrived at **the** optimal solution. Automation bias (Cummings, 2004) – the unwarranted assumption of heightened accuracy of automated systems – can pose a significant confound to interpreting LDA results.

ISER as a Data Source

Although rich, the ISER can also be problematic as a data source. For instance, the ACCJC issues standards the college must comply with to receive accreditation renewal. This means the documents the colleges write are semi-structured, following the structure of the accreditation standard. One wonders what the colleges would produce if they did not have to address specific topics outlined by the accrediting agency. This introduces the risk that some potential elements of identity identified in this study correspond to the identity of the accreditor rather than the college; colleges discussed a specific topic due to the accrediting body's standards, not because it is important to the college.

This also raises a related issue regarding the language of the standards themselves. All the colleges put some quotations from the ACCJC standards in their ISERs. Many colleges used text, but some introduced it as images as well. This language was not specifically excluded during the preprocessing phase, although it was still subject to the same preprocessing adjustments the rest of the texts were subject to. It is unknown what the impact is on the results of this uniform text.

Additionally, it is possible that the ISER does not speak for the entire college but just the dominant coalition that composes the ISER (Foreman & Whetten, 2016, p. 133). Even when all members of an organization are invited to participate, the same people tend to show up repeatedly. These individuals are more active in running the organization, and their voices may carry more weight. This means the latent themes may be more representative of a subset of the college employee population than of the identity of the college as a whole.

COVID-19

It is impossible to write about organizations independently of the impact of the COVID-19 pandemic over the last several years. Many journal articles have yet to be written on the intersection of the pandemic and social processes. So, I write these words with trepidation; we are in uncharted territory. However, everything I have discussed in this dissertation leads me to believe that COVID has had an effect not only on the production of the ISERs but also on the identity of the organizations themselves. For example, my analysis may include some ISERs produced pre-COVID and some ISERs produced during the COVID quarantine. These two groups of documents may not be entirely compatible.

All of these limitations should be kept in mind when interpreting this work. However, despite these limitations, I believe that the findings of the study can be used to improve the design and implementation of future work.

Implications for Practice & Policy

The knowledge gained from this study may help college leaders and change agents see important elements of the deep cultural context surrounding their colleges and secure new tools for educational researchers to observe social phenomena. This research also provides insights into how automated analytic techniques could augment traditional qualitative approaches. Furthermore, the machine learning method explored here may be especially useful for expanding the reach of traditional qualitative analysis as it can serve as a tool for examining large datasets. Finally, understanding the elements a college values can provide new information for decision-making or the construction of instruments to illuminate aspects of the college culture.

The automated method utilized in this study allows the creation of a macroscope (Graham et al., 2016, p. 1) to examine more extensive data sets. Examining the 25 colleges in this study's sample was a daunting but not insurmountable task. The amount of reading required for each college was relatively modest; however, reading the entire ISER for each college, or reading ISERS for a broader sample such as all community colleges nationwide, would be formidable. This study reinforces the idea of the LDA algorithm as a novel tool to examine heretofore unwieldy data sets. It is likely that the California Community College Chancellors' Office could use this tool to see key themes of the entire California Community College system or the ACCJC to compare the impact of standards changes on expressions of valued concepts over time.

The method used in the study has provided insight into how the growing application of automated research tools may impact research findings (Foreman & Whetten, 2016). This study does suggest, to paraphrase Marshall McLuhan, that the method massages the message.

I would not claim this study found college identity themes; the strongest claim I would make is that I was able to extract valued ideas from various colleges. However, this is still very helpful because it allows us to understand better the organization and what themes are important to it. This may facilitate more reflective leadership. Furthermore, the nature of this research may allow a college to document the implementation of those ideals it believes are important. Studies suggest colleges may be more adaptable and resistant to environmental change by understanding their organizational identity (Gioia et al., 2000; Ran & Golden, 2011); it also seems reasonable to assume that understanding a college's values could provide this benefit as well - something the academy may find valuable in these dynamic times (Gioia et al., 2000).

One implication of this work is that it suggests automated methods will not eclipse manual text extraction methods. Rather the two can be employed together to enjoy synergistic effects. This is especially the case if LDA readily identifies latent processes contained within the ISER (as I suspect), as humans may not as easily identify these.

Finally, tools refined in the study may help change agents deliver better interventions targeted toward college identities. The literature suggests that college identities are not uniform and that they cluster (Foreman & Whetten, 2016). Consequently, it is reasonable to assume that different management interventions will be more or less effective when applied to colleges exhibiting specific latent topic clusters. The very nature of the latent themes identified here still raises a key management question: Do I need to work around the college's identity (or values), or do I need to change them? If a revision of a planned intervention is called for, how should that be done? On the other hand, if the change agent needs to change the college's OI (or values), what new themes should it be moving towards? This research can begin to provide some guidance in answering these questions.

Implications for Methods

Several methodological lessons learned emerged from conducting this study as well:

Preprocessing is Complex

Researchers interested in applying techniques such as LDA need to realize that a significant amount of pre-processing work is required to achieve meaningful results. For example, in this project, the context was always a concern. The question "Would the context surrounding the terms in the corpus show through the analysis?" was always present when examining the data. Once the analysis started of both the automated and machine learning methods, it was clear there were two kinds of context one needed to focus on. The first is the context surrounding the theme - which theme is this? A term in this context refers to one theme, but in another context refers to a different theme. Consider the word 'that' in the following two sentences. What did the faculty do about that? What did the college do about that? In both sentences, the entity 'that' refers to would be inferred from the surrounding sentences. What would not be inferred is the entity – either the faculty or the college. When researchers discuss context, I believe they often refer to this type of context.

The second type of context is the context surrounding whom the theme belongs to - what entity are we talking about regarding this theme? I noticed humans fill in this contextual blank well. This became important in this project because the colleges often referred to themselves by their name, i.e., "Folsom Lake College believes ..." This was an issue because using the college name created a smaller pool of co-occurring terms. If every school uses its name when discussing equity (for example), then equity as a latent theme may not be selected by the LDA algorithm because each college has different co-occurring terms with it - their name + "equity." This greatly dilutes the strength of equity as a latent topic (in this example) and makes the terms less likely to emerge as a latent topic. My substitution of "the_college" in place of college names was a critical decision to address this concern.

Researchers need to be vigilant also, as preprocessing can create artifacts. In this study, the preprocessing process created the terms "ref " and "flc_ref," and I could not determine where exactly they came from. This is like the issue of stemming and LDA. Stemming reduces words to their base root term. It allows the aggregation of similar words (say, differing tenses of the same word, for example). Theoretically, this should make it more likely to have topics emerge that use those terms (the tenses no longer dilute the co-occurrence strength). Yet, stemming was not employed in this project because terms like "universal" and "university" reduce to the same stem ("univers"). The stem appears in the list of co-occurring terms, boosting the chances of appearing within a topic but also making it impossible for humans to divine the concept the term refers to. In this study, a similar obfuscation situation arose; some terms were reduced to "ref" and "flc_ref" in the preprocessing phase, and I could not reconstruct where the terms came from.

Term Selection Process

The issues discussed in the previous section suggest that the ISER is best used where the research question can focus on a specific standard. This allows both automated and manual methods to examine data with a high likelihood of containing the topics of interest. This is important for LDA because the algorithm cannot be directed like a human researcher can, and initial algorithm parameter settings can result in different topics. This confirms what the prior literature (Asmussen & Møller, 2019; Chang et al., 2009a; Hong & Davison, 2010) has indicated. This also extends the boundaries of existing research by suggesting the richness of publicly available ISERs as data sources.

The LDA algorithm will generate many more terms for a specific topic than is useful in creating a label for that topic. Many terms will have little impact on defining the concept – like non-coding or junk DNA in relation to the DNA in our genes. The algorithm detects a co-occurrence relationship, but not necessarily a distinctive relationship. Researchers may wish to consider how exactly they will distinguish terms relevant to any given latent topic.

Semi-Structured Texts are Very Different from Unstructured Texts

Many common reference materials for techniques such as LDA use unstructured texts as test datasets. In documents like those, it is reasonable to assume themes are distributed throughout and to extract these themes, one must process the entire text. This is not the case with semi-structured texts. The semi-structured nature of these documents means that some themes are only found in certain sections and missing completely in others. This was an issue for this project because identity themes were only found in one specific section of the ISER, section

I.A.1, to be precise. In hindsight, this was helpful for the dissertation because it illuminated the value themes and made them more prominent, creating a greater contrast with the LDA themes. Practitioners need to be aware of this so they can construct their corpus from exactly the part of the ISER they need.

Another issue may result from comparing sections of text of different sizes. I suspect themes within larger text sections have a greater chance of emerging from analysis because of more co-occurring terms. The ACCJC provides guidelines on how large each section is likely to be, and the language of the standard scaffolds the colleges' responses. The standard implies a question the colleges answer. The discussion within any specific section revolves around the standard, so larger sections may have more co-occurring terms because they relate to the same ideas. This is not the case with unstructured text; with these, the author(s) say what they need to and then move on. The upshot is that researchers need to pay attention to semi-structured texts with sections of greatly different sizes. It is possible that the latent themes LDA extracts will come from the largest chunks of text, and topics from the smaller sections will be missed.

Researcher Input is Still Needed in "Automated" Methods

One unexpected discovery of the LDA algorithm was that a surprisingly large amount of researcher interpretation was required. For example, selecting the terms to include within the various latent topics was not as straightforward as it first appeared. Do you choose the terms with the largest posterior probability, the largest lift, or the largest topic composition (the term appears in multiple topics but predominates in just one)? Similar questions arise concerning the number of topics the algorithm should search for, the way the data is preprocessed, and even what documents (or parts of documents) should go into the corpus. The bottom line is that while LDA may look quantitative, researchers need to be aware that significant human judgment is still required.

Recommendations

This study explored the contrasts and similarities in latent Organizational Identity topics extracted from Institutional Self Evaluation Reports for community colleges by automated topic extraction and manual content analysis methods. Both methods extracted latent themes from the ISERs – but they were not the same latent themes. I offer the following recommendations to educational practitioners interested in using these tools to understand better their environment.

Identification of valued themes was straightforward for human content analysis. Value statements in the ISER made understanding important motivational ideas much easier. This should make it easier for a change agent to understand the motivations of a local college. However, the resistance of these themes to LDA detection may make it harder for entities like the Chancellor's Office to understand this aspect of the behavior of the entire community college system. Additional work should be conducted to extend this study to all the California Community Colleges to understand the impact of these processes at the system level.

Applying LDA to the ISER may result in developing a baseline of common processes for given accreditation or administrative activities. In this study, for example, the algorithm appeared to extract latent topics related to processes involving the implementation of the college's mission. If this is indeed what has happened, this information would be extremely valuable for administrators and accreditors to know as it could help them identify strengths and weaknesses in the colleges' processes. This would be doubly so if there were a link between these processes and outcomes. Furthermore, each area (student services, faculty, etc.) may have its own set of processes. I suggest further research in this area is needed.

ISER Standardization Facilitates Access

Accrediting agencies may wish to give some thought to changing the format of the ISER to make the document more machine-readable. The general idea would be to make the ISER a more uniform document by standardizing ISER production (for example, "Do not include specification language in your ISER" or "Do not include pictures in the ISER submitted to the

ACCJC.”). Colleges could produce two versions of their ISER with the same content, one for public consumption and appearing similar to what is produced today and another similar to legal documents (very uniform, utilitarian documents) for submission to the accreditor. Yes, the cost of producing an ISER would increase, but it would be much easier for all stakeholders to access critical information. This, in turn, may allow administrators deep insight into the processes at their schools or the ACCJC to reduce the burden of site visits. An extreme alternative would be for the ACCJC to decide the colleges should submit a very austere, text only ISER, and colleges could pocket the resulting savings from no longer necessary photography or typesetting services. In general, policy leaders may wish to devote some thought to making the documents they require more machine-readable by making them more standardized. This would assist practitioners in utilizing automated methods on the data.

Finally, I recommend that researchers consider using automated text extraction algorithms like LDA. These can provide a unique and different lens to view the data. Furthermore, the ISER offers an unusual dataset to examine the activities occurring at any specific college. Educational research has a place for algorithms like LDA and datasets like ISERs.

Conclusions and Future Research

This study explored the contrasts and similarities in latent Organizational Identity topics extracted from Institutional Self Evaluation Reports for community colleges by automated topic extraction and manual content analysis methods. Both methods extracted latent themes from the ISERs – but they were not the same latent themes. The automated method extracted process latent topics from the corpus. The manual method extracted value themes, but it cannot be determined from the corpus if those value themes rise to the level of organizational identities. Several of the findings were consistent with the literature. This work also extends the academic conversations by providing researchers and practitioners some clarity about what to expect when automated and manual methods are used together and some of the ramifications of using ISERs as a dataset.

The relationship between organizational identity and organizational performance needs to be clarified. There is some evidence in the literature to suggest a relationship between organizational identity and organizational performance. Furthermore, this study suggests that LDA can identify key institutional processes by examining the ISER. However, this relationship with respect to college performance is complex and not fully understood. Future research could more deeply explore the relationship between organizational identity themes and organizational performance.

There is potential in automated topic extraction techniques like LDA. This study has shown that LDA can be a valuable tool for identifying latent themes in college ISERs. However, it is important to be aware of the limitations of this approach, such as it can be challenging to interpret the results and that it may not be able to capture relationships inherent in the nuance of human language. Further research could explore the potential of automated topic extraction methods like LDA for other types of community college research.

This project succeeded in its goal of generating new information for college change agents to use in planning interventions. Topic extraction methods have the potential for community college research. This study has shown that LDA can be a valuable tool in identifying latent process themes in college ISERs. At the same time, manual semantic analysis was helpful in understanding latent topics the colleges valued. We know our changing world poses ever-present management challenges for community colleges. This research may offer tools to assist change agents in increasing their chances of success with their change interventions.

REFERENCES

AACRAO. (2018). *The rapidly changing nature of community college*. The American Association of Collegiate Registrars and Admissions Officers. Retrieved 1/21/2023 from https://www.aacrao.org/resources/newsletters-blogs/aacrao-connect/article/the-rapidly-changing-nature-of-community-college

Accrediting Commission for Community and Junior Colleges. (2022). *Guide to Institution Self-Evaluation, Improvement, and Peer Review*. https://accjc.org/wp-content/uploads/Guide-to-Institutional-Self-Evaluation-Improvement-Peer-Review_Jan2020.pdf

Albert, S., & Whetten, D. (1985). Organizational Identity. *Research in Organizational Behavior*, *7*, 263 - 295.

AlSumait, L., Barbará, D., Gentle, J., & Domeniconi, C. (2009). Topic significance ranking of LDA generative models. Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part I 20,

American River College. (2021). Institutional Self Evaluation Report.

Anthony, C., & Tripsas, M. (2016). Organizational identity and innovation. In M. G. Pratt, M. Schultz, B. E. Ashforth, & D. Ravasi (Eds.), *The Oxford handbook of organizational identity* (Vol. 1, pp. 417-435). Oxford University Press.

Argyris, C., & Schön, D. A. (1996). *Organizational Learning II: Theory, Method, and Practice*. Addison-Wesley.

Asmussen, C. B., & Møller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data*, *6*(1), 1-18.

Ayers, D. F. (2002). Mission priorities of community colleges in the southern United States. *Community College Review*, *30*(3), 11-30.

Ayers, D. F. (2015). Credentialing structures, pedagogies, practices, and curriculum goals: Trajectories of change in community college mission statements. *Community College Review*, *43*(2), 191-214.

Basko, A. (2022). *Stop Playing It Safe: The Peril of the Generic College*. The Chronicle of Higher Education. Retrieved 2/20/2023 from https://www.chronicle.com/article/stop-playing-it-safe-the-peril-of-the-generic-college

Bayrak, T. (2020). A content analysis of top-ranked universities' mission statements from five global regions. *International Journal of Educational Development*, *72*, 102130.

Berg, B. L., & Lune, H. (2007). *Qualitative research methods for the social sciences* (6 ed.). Pearson.

Birnbaum, R., & Snowdon, K. (2003). Management fads in higher education. *The Canadian Journal of Higher Education*, *33*(2).

Bischof, J., & Airoldi, E. M. (2012). Summarizing topical content with word frequency and exclusivity. Proceedings of the 29th international conference on machine learning (icml-12),

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.

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

Borromeo, R. M., & Toyama, M. (2015). Automatic vs. crowdsourced sentiment analysis. Proceedings of the 19th International Database Engineering & Applications Symposium,

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, *15*(5), 662-679.

Brandall, b. (2018). *How to Write Your Company Mission Statement: 200 Top Examples*

Analyzed. process.st. Retrieved 11/24/22 from https://www.process.st/company-mission-statement/

Braun, V., & Clarke, V. (2022). *Thematic analysis: A practical guide*. SAGE Publications. Kindle

Buenano-Fernandez, D., González, M., Gil, D., & Luján-Mora, S. (2020). Text mining of open-ended questions in self-assessment of university teachers: An LDA topic modeling approach. *IEEE Access*, *8*, 35318-35330.

Burnett, C. A. (2020). Diversity under Review: HBCUs and Regional Accreditation Actions. *Innovative Higher Education*, *45*(1), 3-15. https://doi.org/10.1007/s10755-019-09482-w

Butte-Glenn Community College District. (2021). *2021 Institutional Self Evaluation Report*. https://www.butte.edu/departments/governance/accreditation/2021-iser.html

California Community College Chancellor's Office. (2023). *An Alphabetical Listing of California Community Colleges*. Retrieved 3/24/23 from https://www.cccco.edu/Students/Find-a-College/College-Alphabetical-Listing

California State Auditor. (2013). *California Community College Accreditation: Colleges are Treated Inconsistently and Opportunities Exist for Improvement in the Accreditation Process*. www.auditor.ca.gov

Chait, R. (1979). Mission madness strikes our colleges. *Chronicle of Higher Education*, *18*(36), A36.

Chaney, A., & Blei, D. (2012). Visualizing topic models. Proceedings of the International AAAI Conference on Web and Social Media,

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009a). Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, *22*.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009b). Reading tea leaves: How humans interpret topic models. Advances in neural information processing systems,

Chen, X., Zou, D., Cheng, G., & Xie, H. (2020). Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of Computers & Education. *Computers & Education*, *151*, 103855.

Chuang, J., Gupta, S., Manning, C., & Heer, J. (2013). Topic model diagnostics: Assessing domain relevance via topical alignment. International conference on machine learning,

Clark, S. M., Gioia, D. A., Ketchen Jr, D. J., & Thomas, J. B. (2010). Transitional identity as a facilitator of organizational identity change during a merger. *Administrative Science Quarterly*, *55*(3), 397-438.

Consumnes River College. (2021). Institutional Self Evaluation Report.

Conway, M. (2006). The subjective precision of computers: A methodological comparison with human coding in content analysis. *Journalism & Mass Communication Quarterly*, *83*(1), 186-200.

Corley, K. G., & Gioia, D. A. (2004). Identity ambiguity and change in the wake of a corporate spin-off. *Administrative Science Quarterly*, *49*(2), 173-208.

Cortés Sánchez, J. D. (2018). Mission statements of universities worldwide: Text mining and visualization. *Intangible Capital*, *14*(4), 584-603.

Cortés-Sánchez, J. (2017). Mission and vision statements of universities worldwide: A content analysis. *Documentos De Investigación, Facultad de Administración*(152), 2463-1892.

Creswell, J. W., & Creswell, J. D. (2020). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th, Ed.). Sage publications.

Cummings, M. (2004). Automation bias in intelligent time critical decision support systems. AIAA 1st intelligent systems technical conference,

Cypress College. (2017). 2017 Institutional Self Evaluation Report in Support of Reaffirmation of Accreditation.

Davies, S. W., & Glaister, K. W. (1996). Spurs to higher things? Mission statements of UK universities. *Higher Education Quarterly*, *50*(4), 261-294.

De Graaf, R., & van der Vossen, R. (2013). Bits versus brains in content analysis. Comparing the advantages and disadvantages of manual and automated methods for content analysis. *38*(4), 433-443. https://doi.org/10.1515/commun-2013-0025

Dedoose. (2023). *dedoose*. In https://www.dedoose.com/

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American sociological review*, 147-160.

Djeukeng, B. N. (2014). *An exploration of compliance predictors of the institutional effectiveness requirements of the Southern Association of Colleges and Schools Commission on Colleges' baccalaureate instittutions between 2008 and 2012* William & Mary].

Drucker, P. F. (1974). Tasks, responsibilities, practices. *NY: Truman*.

Dutton, J. E., Dukerich, J. M., & Harquail, C. V. (1994). Organizational images and member identification. *Administrative Science Quarterly*, 239-263.

Education, U. S. D. o. (2022). *College Accreditation in the United States*. Retrieved 10/27/2022 from https://www2.ed.gov/admins/finaid/accred/index.html

Efe, I., & Ozer, O. (2015). A corpus-based discourse analysis of the vision and mission statements of universities in Turkey. *Higher Education Research & Development*, *34*(6), 1110-1122.

Fiol, C. M. (1991). Managing culture as a competitive resource: An identity-based view of sustainable competitive advantage. *Journal of Management*, *17*(1), 191-211.

Fiol, C. M. (2002). Capitalizing on paradox: The role of language in transforming organizational identities. *Organization Science*, *13*(6), 653-666.

Firmin, M. W., & Gilson, K. M. (2009). Mission statement analysis of CCCU member institutions. *Christian Higher Education*, *9*(1), 60-70.

Firth, J. R. (1957). *Studies in Linguistic Analysis*. Wiley-Blackwell.

Foreman, P., & Whetten, D. (2016). Measuring Organizational Identity: takinig stock and looking forward. In M. G. Pratt, M. Schultz, B. E. Ashforth, & D. Ravasi (Eds.), *The Oxford handbook of organizational identity*. Oxford University Press.

Fox, S., Scheffler, I., & Marom, D. (2003). *Visions of Jewish education*. Cambridge University Press.

Fresno City College. (2018). Institutional Self Evaluation Report of Educational Quality and Institutional Effectiveness in Support of Reaffirmation of Accreditation.

Gándara, D., & Daenekindt, S. (2022). Accountability or Equity: Combining Topic Models and Qualitative Analysis to Examine Public Rhetoric about Performance-Based Funding. *Educational evaluation and policy analysis*, *44*(4), 734-758.

Gardner, M. J., Lutes, J., Lund, J., Hansen, J., Walker, D., Ringger, E., & Seppi, K. (2010). The topic browser: An interactive tool for browsing topic models. Nips workshop on challenges of data visualization,

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Gil, W.-J., Kim, J.-W., Park, K.-R., & Cho, H.-J. (2021). An Analysis of Research Trends in AI Education based on LDA. *Review of International Geographical Education Online*, *11*(2), 254-262.

Gioia, D. A., & Hamilton, A. L. (2016). Great Debates in Organizational Identity Study. In M. G. Pratt, M. Schultz, B. E. Ashforth, & D. Ravasi (Eds.), *The Oxford handbook of organizational identity* (pp. 21 - 38). Oxford University Press.

Gioia, D. A., Schultz, M., & Corley, K. G. (2000). Organizational Identity, Image, and Adaptive Instability. *Academy of Management Review*, *25*, 63-81.

Glynn, M. A. (2000). When cymbals become symbols: Conflict over organizational identity within a symphony orchestra. *Organization Science*, *11*(3), 285-298.

Graham, S., Milligan, I., Weingart, S. B., & Martin, K. (2016). *Exploring big historical data: the historian's macroscope*. World Scientific.

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

Hartnell College. (2019). Institutional Self Evaluation Report.

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in twitter. Proceedings of the first workshop on social media analytics,

Hossain, A. D., Hossain, A. R., & Kouar, M. (2019). Optimizing Assessment Tasks for Institutional and Program-Level Accreditations: A Case Study of Accreditation Requirements of MSCHE and ABET. *Journal of Assessment and Institutional Effectiveness*, *9*(1-2), 96-120.

Ignatow, G., & Mihalcea, R. (2017). *Text mining : a guidebook for the social sciences*. SAGE. https://search.library.ucdavis.edu/permalink/f/12qmtm2/01UCD_ALMA21283728000003126

Ignatow, G., & Mihalcea, R. (2018). *An introduction to text mining : research design, data collection, and analysis*. SAGE.

Irvine Valley College. (2017). *2017 Institutional Self Evaluation Report*. https://www.ivc.edu/files/accreditation/pdf/IVC_2017_Self_Evaluation.pdf

Kezar, A. (2018). *How colleges change: Understanding, leading, and enacting change*. Routledge.

Kirilenko, A. P., Stepchenkova, S. O., Kim, H., & Li, X. (2018). Automated sentiment analysis in tourism: Comparison of approaches. *Journal of Travel Research*, *57*(8), 1012-1025.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Kuckartz, U. (2019). Qualitative Text Analysis: A Systematic Approach. In G. Kaiser & N. Presmeg (Eds.), *Compendium for Early Career Researchers in Mathematics Education*. Springer Nature.

Kuenssberg, S. (2011). The discourse of self-presentation in Scottish university mission statements. *Quality in higher education*, *17*(3), 279-298.

Lambert, B. (2018). *A student's guide to Bayesian statistics*. SAGE Publications.

Laney College. (2020). Institutional Self Evaluation Report.

Las Positas College. (2022). 2022 Institutional Self Evaluation Report. https://www.laspositascollege.edu/accreditation/

Levin, J. (2001). *Globalizing the community college: Strategies for change in the twenty-first century*. Springer.

Lewis, J. (2017). With deep learning, the data-rich get richer. *InfoWorld*. Retrieved 2/20/2023, from https://www.infoworld.com/article/3181736/with-deep-learning-the-data-rich-get-richer.html

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of broadcasting & electronic media*, *57*(1), 34-52.

Linderman, A. (2001). Computer content analysis and manual coding techniques: A comparative analysis. *Progress in Communication Sciences*, 97-110.

Matsuda, Y., Sekiya, T., & Yamaguchi, K. (2018). Curriculum analysis of computer science departments by simplified, supervised LDA. *Journal of Information Processing*, *26*, 497-508.

Merriam, S. B., & Tisdell, E. J. (2016). *Qualitative research : a guide to design and implementation* (Fourth edition. ed.). Jossey-Bass.

Mertz, D. (2021). *Cleaning data for effective data science*. Packt Publishing.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. Proceedings of the 2011 conference on empirical methods in natural language processing,

MiraCosta College. (2016). *Institutional Self Evaluation Report In Support of Reaffirmation of Accreditation*.

Mitchell, T. M., & Mitchell, T. M. (1997). *Machine learning* (Vol. 1). McGraw-hill New York.

Modesto Junior College. (2017). Institutional Self Evaluation Report.

Morphew, C. C., & Hartley, M. (2006). Mission statements: A thematic analysis of rhetoric across institutional type. *The Journal of Higher Education*, *77*(3), 456-471.

Muller, A. E., Ames, H. M. R., Jardim, P. S. J., & Rose, C. J. (2022). Machine learning in systematic reviews: Comparing automated text clustering with Lingo3G and human researcher categorization in a rapid review. *Research Synthesis Methods*, *13*(2), 229-241.

Nacos, B. L., Shapiro, R. Y., Young, J. T., Fan, D. P., Kjellstrand, T., & McCaa, C. (1991). Content analysis of news reports: Comparing human coding and a computer-assisted method. *Communication (New York)*, *12*(2), 111-128.

Nag, R., Corley, K. G., & Gioia, D. A. (2007). The intersection of organizational identity, knowledge, and practice: Attempting strategic change via knowledge grafting. *Academy of Management Journal*, *50*(4), 821-847.

Newsom, W., & Hayes, C. (1991). Are Mission Statements Worthwhile? *Planning for Higher Education*, *19*(2), 28-30.

Oxford Languages. (2023). *Semantic, defiinition*. Google English Dictionary. Retrieved 03/21/2023 from https://www.google.com/search?q=semantics&oq=semantic&aqs=chrome.0.0i131i433i512j69i57j0i131i433i512l2j46i433i512j0i131i433i512l3j0i433i512j0i131i433i512.2261j0j7&sourceid=chrome&ie=UTF-8

Palmer, T. B., & Short, J. C. (2008). Mission statements in US colleges of business: An empirical examination of their content with linkages to configurations and performance. *Academy of Management Learning & Education*, *7*(4), 454-470.

Palomar College. (2022). Institutional Self Evaluation Report. https://www.palomar.edu/accreditation/

Papadimitriou, C. H., Tamaki, H., Raghavan, P., & Vempala, S. (1998). Latent semantic indexing: A probabilistic analysis. Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems,

Pekarsky, D. (2007). Vision and education: Arguments, counterarguments, rejoinders. *American Journal of Education*, *113*(3), 423-450.

Perez-Encinas, A., Rodriguez-Pomeda, J., & de Wit, H. (2021). Factors influencing student mobility: a comparative European study. *Studies in Higher Education*, *46*(12), 2528-2541.

Peters, T. J., & Waterman, R. H. (1982). *In search of excellence: Lessons from America's best-run companies*. New York: Harper & Row.

Petriglieri, J. L., & Devine, B. A. (2016). Mobilizing organizational action against identity threats. In M. G. Pratt, M. Schultz, B. E. Ashforth, & D. Ravasi (Eds.), *The Oxford handbook of organizational identity* (pp. 239-256). Oxford University Press.

Pratt, M. G. (2000). The good, the bad, and the ambivalent: Managing identification among Amway distributors. *Administrative Science Quarterly*, *45*(3), 456-493.

Ramage, D., Rosen, E., Chuang, J., Manning, C. D., & McFarland, D. A. (2009). Topic modeling for the social sciences. NIPS 2009 workshop on applications for topic models: text and beyond,

Ramesh, A., Goldwasser, D., Huang, B., Daumé III, H., & Getoor, L. (2014). Understanding MOOC discussion forums using seeded LDA. Proceedings of the ninth workshop on innovative use of NLP for building educational applications,

Ran, B., & Golden, T. J. (2011). Who are we? The social construction of organizational identity through sense-exchanging. *Administration & Society*, *43*(4), 417-445.

Rosenberg, S. D., Schnurr, P. P., & Oxman, T. E. (1990). Content analysis: A comparison of manual and computerized systems. *Journal of personality assessment*, *54*(1-2), 298-310.

Ruef, M., & Nag, M. (2015). *The Classification of Organizational Forms Theory and Application to the Field of Higher Education*. Stanford University Press.

Rybinski, K. (2020). Are rankings and accreditation related? Examining the dynamics of higher education in Poland. *Quality Assurance in Education*. https://www.emerald.com/insight/content/doi/10.1108/QAE-03-2020-0032/full/html

Saldaña, J. (2021). *The coding manual for qualitative researchers*.

Seeber, M., Barberio, V., Huisman, J., & Mampaey, J. (2019). Factors affecting the content of universities' mission statements: an analysis of the United Kingdom higher education system. *Studies in Higher Education*, *44*(2), 230-244.

Sequoias Community College District. (2018). *Institutional Self Evaluation Report*.

Serafin, M. J. (2014). *Accreditation Follow-Up: A Grounded Theory Qualitative Study of WASC-Accredited Private Schools in Southern California* (Publication Number 3581359) [Ed.D., La Sierra University]. ProQuest Dissertations & Theses A&I. Ann Arbor. https://www.proquest.com/docview/1552469431?accountid=14505

https://ucdavis-primo.hosted.exlibrisgroup.com/openurl/01UCD/01UCD_SP?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&genre=dissertations+%26+theses&sid=ProQ:ProQuest+Dissertations+%26+Theses+A%26I&atitle=&title=Accreditation+follow-up%3A+A+grounded+theory+qualitative+study+of+WASC-accredited+private+schools+in+Southern+California&issn=&date=2014-01-01&volume=&issue=&spage=&au=Serafin%2C+Marsha+Jean&isbn=978-1-321-13138-3&jtitle=&btitle=&rft_id=info:eric/&rft_id=info:doi/

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. Proceedings of the workshop on interactive language learning, visualization, and interfaces,

Silge, J., & Robinson, D. (2017). *Text mining with R : a tidy approach* (First edition ed.). O'Reilly Media. https://search.library.ucdavis.edu/permalink/f/12qmtm2/01UCD_ALMA51358055080003126

Simon, A. F. (2001). A unified method for analyzing media framing. *Communication in US elections: New agendas*, 75-89.

Sjøvaag, H., Moe, H., & Stavelin, E. (2012). Public service news on the Web: A large-scale content analysis of the Norwegian Broadcasting Corporation's online news. *Journalism Studies*, *13*(1), 90-106. https://doi.org/https://doi.org/10.1080/1461670X.2011.578940

Sjøvaag, H., & Stavelin, E. (2012). Web media and the quantitative content analysis: Methodological challenges in measuring online news content. *Convergence*, *18*(2), 215-229.

Sodhi, R. (2016). *Accrediting processes and institutional effectiveness at a California community college* pdf

Soysal, Y. N., & Baltaru, R.-D. (2021). University as the producer of knowledge, and economic and societal value: the 20th and twenty-first century transformations of the UK higher education system. *European Journal of Higher Education*, *11*(3), 312-328.

Straumsheim, C. (2017). Trackinig the Evolution of Student Success. *Inside Higher Education*. https://www.insidehighered.com/quicktakes/2017/02/06/tracking-evolution-student-success

Sun, M., Liu, J., Zhu, J., & LeClair, Z. (2019). Using a Text-as-Data Approach to Understand Reform Processes: A Deep Exploration of School Improvement Strategies. *Educational evaluation and policy analysis*, *41*, 510 - 536.

Taddy, M. (2012). On estimation and selection for topic models. Artificial intelligence and statistics,

Theule, R. W. (2012). *An Exploratory, Quantitative Study of Accreditation Actions Taken by the Western Association of Schools and Colleges' Accrediting Commission for Community and Junior Colleges (WASC-ACCJC) since 2002* ERIC. https://search.proquest.com/docview/1651851680?accountid=14505

https://ucdavis-primo.hosted.exlibrisgroup.com/openurl/01UCD/01UCD_SP?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&genre=dissertations+%26+theses&sid=ProQ:ERIC&atitle=&title=An+Exploratory%2C+Quantitative+Study+of+Accreditation+Actions+Taken+by+the+Western+Association+of+Schools+and+Colleges%27+Accrediting+Commission+for+Community+and+Junior+Colleges+%28WASC-ACCJC%29+since+2002&issn=&date=2012-01-01&volume=&issue=&spage=&au=Theule%2C+Ryan+William&isbn=9781267451101&jtitle=&btitle=&rft_id=info:eric/ED547562&rft_id=info:doi/

Trainor, L. R., & Bundon, A. (2021). Developing the craft: Reflexive accounts of doing reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *13*(5), 705-726. https://doi.org/10.1080/2159676X.2020.1840423

Tsang, S. (2020). *Best Mission Statements: 12 Examples You Need to See*. Retrieved 11/24/22 from https://www.fond.co/blog/best-mission-statements/

Wallach, H., Mimno, D., & McCallum, A. (2009). Rethinking LDA: Why priors matter. *Advances in neural information processing systems*, *22*.

Wang, J., Gibson, A. M., Salinas, L., Solis, F., & Slate, J. R. (2007). Thematic Differences in Mission Statements between Four-Year Public Institutions and Two-Year Colleges in Texas. *International Electronic Journal for Leadership in Learning*, *11*, 1.

Weick, K. E. (2001). *Making sense of the organization*. Blackwell Publishers.

West Los Angeles College. (2023). *Institutional Self-Evaluation Report*. https://studentlaccd.sharepoint.com/teams/LACCDDWAccreditation/WLAC/Forms/AllItems.aspx?id=%2Fteams%2FLACCDDWAccreditation%2FWLAC%2FWest%5F2023%5FISER%20%2D%20Final%207%5F29%5F22%5FADA%2Epdf&parent=%2Fteams%2FLACCDDWAccreditation%2FWLAC&p=true&ga=1

Whetten, D. A. (2006). Albert and Whetten revisited: Strengthening the concept of organizational identity. *Journal of management inquiry*, *15*(3), 219-234.

Whetten, D. A., & Mackey, A. (2002). A social actor conception of organizational identity and its implications for the study of organizational reputation. *Business & society*, *41*(4), 393-414.

Woodland Community College. (2018). Institutional Self Evaluation Report.

Yin, B., & Yuan, C.-H. (2022). Detecting latent topics and trends in blended learning using LDA topic modeling. *Education and Information Technologies*, 1-24.

APPENDIX A: LIST OF COLLEGES IN SAMPLE

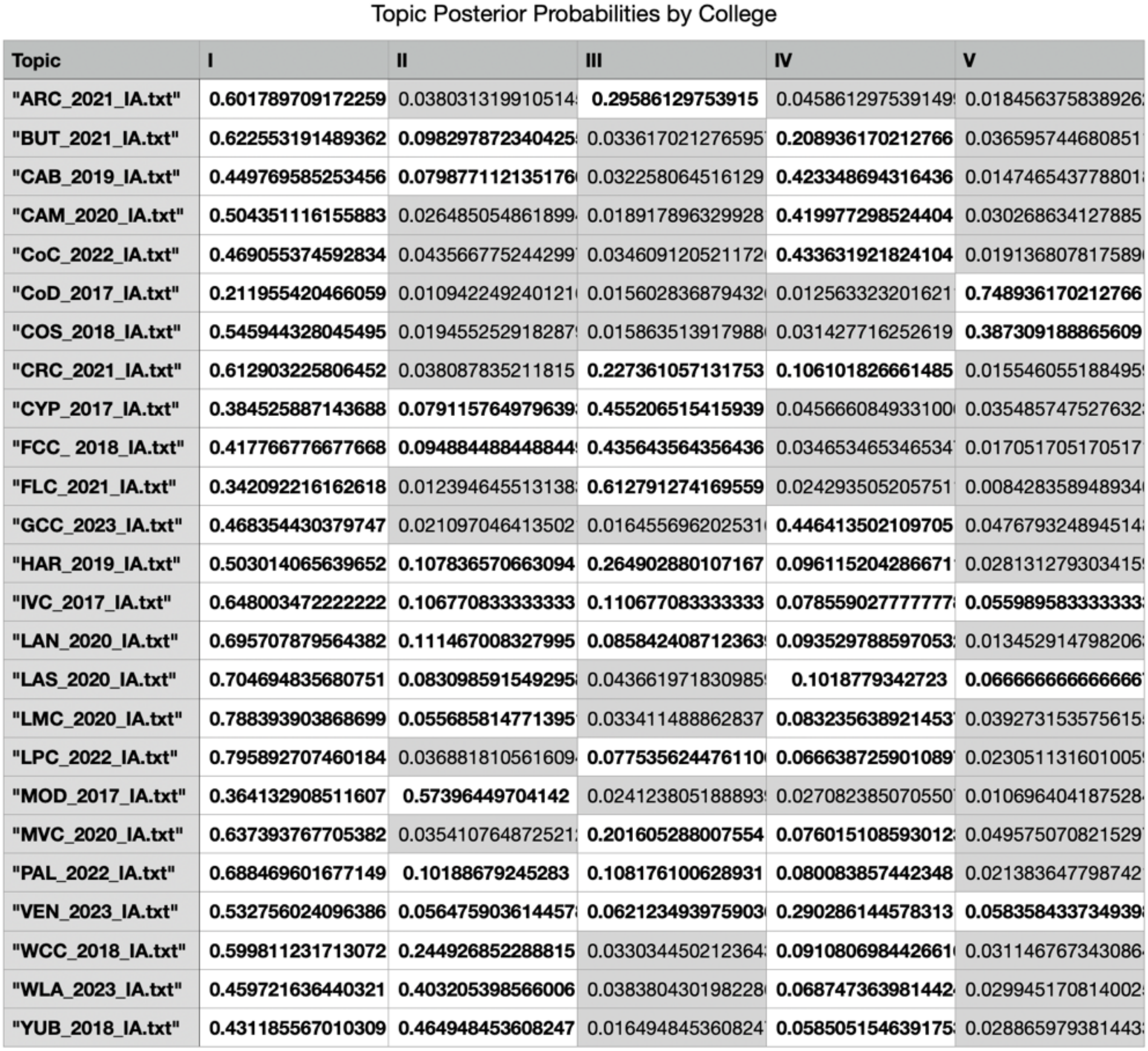
College Name	Filename
American River College	ARC_2021_IA.txt
Butte College	BUT_2021_IA.txt
Cabrillo College	CAB_2019_IA.txt
El Camino College	CAM_2020_IA.txt
College of the Canyons	CoC_2022_IA.txt
College of the Desert	CoD_2017_IA.txt
College of the Sequoias	COS_2018_IA.txt
Cosumnes River College	CRC_2021_IA.txt
Cypress College	CYP_2017_IA.txt
Fresno City College	FCC_ 2018_IA.txt
Folsom Lake College	FLC_2021_IA.txt
Glendale Community College	GCC_2023_IA.txt
Hartnell College	HAR_2019_IA.txt
Irvine Valley College	IVC_2017_IA.txt
Laney College	LAN_2020_IA.txt
Lassen Community College	LAS_2020_IA.txt
Los Medanos College	LMC_2020_IA.txt
Las Positas College	LPC_2022_IA.txt
Modesto Junior College	MOD_2017_IA.txt
Moreno Valley College	MVC_2020_IA.txt
Palomar College	PAL_2022_IA.txt
Ventura College	VEN_2023_IA.txt
Woodland Community College	WCC_2018_IA.txt
West Los Angeles College	WLA_2023_IA.txt
Yuba College	YUB_2018_IA.txt

Filename format: The underscore ('_') character separates key metadata attributes encoded into the file name. This makes it relatively easy for the computer to access the metadata. The first three letters are the college abbreviation – which may or may not be the same abbreviation the college itself uses. The four digit number is the year the ISER was published. Finally, the AI refers to the section of the ISER this file contains (The full documents replace IA with ISER).

APPENDIX B: COLLEGE DISTINCTIVENESS BY POSTERIOR PROBABILITIES

Figure 18

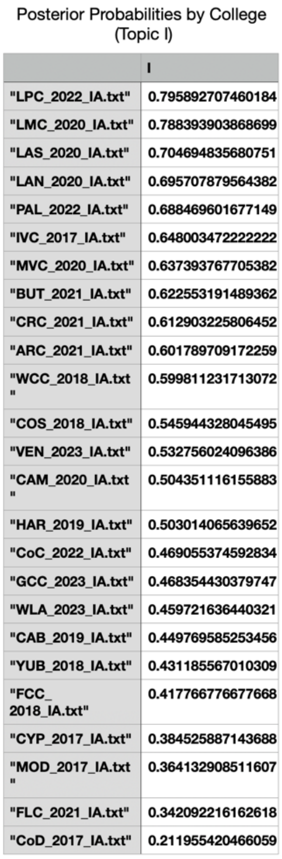
Topic Posterior Probabilities by College



Topic Posterior Probabilities by College

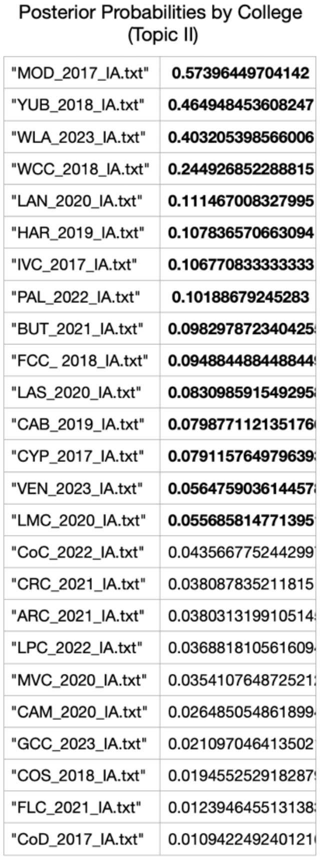
Topic	I	II	III	IV	V
"ARC_2021_IA.txt"	**0.601789709172259**	0.038031319910514	**0.29586129753915**	0.045861297539149	0.018456375838926
"BUT_2021_IA.txt"	**0.622553191489362**	**0.098297872340425**	0.033617021276595	**0.208936170212766**	0.036595744680851
"CAB_2019_IA.txt"	**0.449769585253456**	**0.079877112135176**	0.032258064516129	**0.423348694316436**	0.014746543778801
"CAM_2020_IA.txt"	**0.504351116155883**	0.026485054861899	0.018917896329928	**0.419977298524404**	0.030268634127885
"CoC_2022_IA.txt"	**0.469055374592834**	0.043566775244299	0.034609120521172	**0.433631921824104**	0.019136807817589
"CoD_2017_IA.txt"	**0.211955420466059**	0.010942249240121	0.015602836879432	0.012563323201621	**0.748936170212766**
"COS_2018_IA.txt"	**0.545944328045495**	0.019455252918287	0.015863513917988	0.031427716252619	**0.387309188865609**
"CRC_2021_IA.txt"	**0.612903225806452**	0.038087835211815	**0.227361057131753**	**0.106101826661485**	0.015546055188495
"CYP_2017_IA.txt"	**0.384525887143688**	**0.079115764979639**	**0.455206515415939**	0.045666084933100	0.035485747527632
"FCC_ 2018_IA.txt"	**0.417766776677668**	**0.094884488448844**	**0.435643564356436**	0.034653465346534	0.017051705170517
"FLC_2021_IA.txt"	**0.342092216162618**	0.012394645513138	**0.612791274169559**	0.024293505205751	0.008428358948934
"GCC_2023_IA.txt"	**0.468354430379747**	0.021097046413502	0.016455696202531	**0.446413502109705**	0.047679324894514
"HAR_2019_IA.txt"	**0.503014065639652**	**0.107836570663094**	**0.264902880107167**	**0.096115204286671**	0.028131279303415
"IVC_2017_IA.txt"	**0.648003472222222**	**0.106770833333333**	**0.110677083333333**	**0.078559027777777**	**0.055989583333333**
"LAN_2020_IA.txt"	**0.695707879564382**	**0.111467008327995**	**0.085842408712363**	**0.093529788597053**	0.013452914798206
"LAS_2020_IA.txt"	**0.704694835680751**	**0.083098591549295**	0.043661971830985	**0.1018779342723**	**0.066666666666666**
"LMC_2020_IA.txt"	**0.788393903868699**	**0.055685814771395**	0.033411488862837	**0.083235638921453**	0.039273153575615
"LPC_2022_IA.txt"	**0.795892707460184**	0.036881810561609	**0.077535624476110**	**0.066638725901089**	0.023051131601005
"MOD_2017_IA.txt"	**0.364132908511607**	**0.57396449704142**	0.024123805188893	0.027082385070550	0.010696404187528
"MVC_2020_IA.txt"	**0.637393767705382**	0.035410764872521	**0.201605288007554**	**0.076015108593012**	0.049575070821529
"PAL_2022_IA.txt"	**0.688469601677149**	**0.10188679245283**	**0.108176100628931**	**0.080083857442348**	0.021383647798742
"VEN_2023_IA.txt"	**0.532756024096386**	**0.056475903614457**	**0.062123493975903**	**0.290286144578313**	**0.058358433734939**
"WCC_2018_IA.txt"	**0.599811231713072**	**0.244926852288815**	0.033034450212364	**0.091080698442661**	0.031146767343086
"WLA_2023_IA.txt"	**0.459721636440321**	**0.403205398566006**	0.038380430198228	**0.068747363981442**	0.029945170814002
"YUB_2018_IA.txt"	**0.431185567010309**	**0.464948453608247**	0.016494845360824	**0.058505154639175**	0.028865979381443

This material explores the posterior probability that a specific topic is important to a given college. The posterior probability is the prediction that the college will be associated with a particular topic. I selected 0.05 as a level of significance and bolded all colleges where the posterior probability was equal to or greater than that level. I also selected a light gray background for non-bolded cells to make reading easier. What this says is that this topic is important enough to the college for inclusion within a specific topic. In Figure 18, colleges and their associated posterior probabilities are listed alphabetically. For example, Topic I (whatever that topic is) is important to American River College (and all the other colleges in the sample). Consequently, it is hard to claim that topic one would be distinctive. Topic five, on the other hand, is only significant to five colleges in the sample. Is that a small enough group to designate the topic as distinctive? What about topics two, three, and four? Are they distinctive? There are no guidelines to answer that question. The following are the individual posterior probabilities by the significance level for each topic.



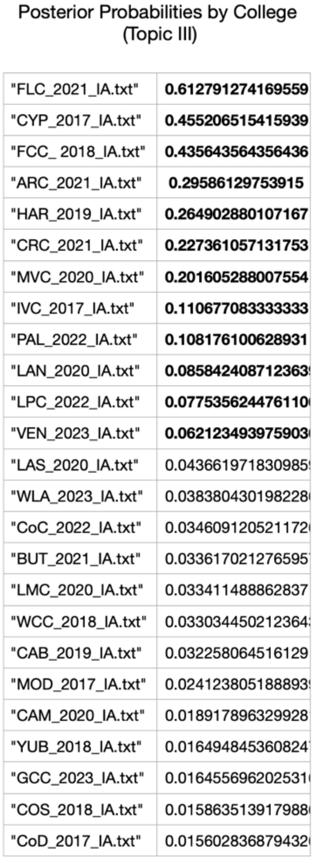
Posterior Probabilities by College
(Topic I)

	I
"LPC_2022_IA.txt"	**0.795892707460184**
"LMC_2020_IA.txt"	**0.788393903868699**
"LAS_2020_IA.txt"	**0.704694835680751**
"LAN_2020_IA.txt"	**0.695707879564382**
"PAL_2022_IA.txt"	**0.688469601677149**
"IVC_2017_IA.txt"	**0.648003472222222**
"MVC_2020_IA.txt"	**0.637393767705382**
"BUT_2021_IA.txt"	**0.622553191489362**
"CRC_2021_IA.txt"	**0.612903225806452**
"ARC_2021_IA.txt"	**0.601789709172259**
"WCC_2018_IA.txt "	**0.599811231713072**
"COS_2018_IA.txt"	**0.545944328045495**
"VEN_2023_IA.txt"	**0.532756024096386**
"CAM_2020_IA.txt "	**0.504351116155883**
"HAR_2019_IA.txt"	**0.503014065639652**
"CoC_2022_IA.txt"	**0.469055374592834**
"GCC_2023_IA.txt"	**0.468354430379747**
"WLA_2023_IA.txt"	**0.459721636440321**
"CAB_2019_IA.txt"	**0.449769585253456**
"YUB_2018_IA.txt"	**0.431185567010309**
"FCC_ 2018_IA.txt"	**0.417766776677668**
"CYP_2017_IA.txt"	**0.384525887143688**
"MOD_2017_IA.txt "	**0.364132908511607**
"FLC_2021_IA.txt"	**0.342092216162618**
"CoD_2017_IA.txt"	**0.211955420466059**



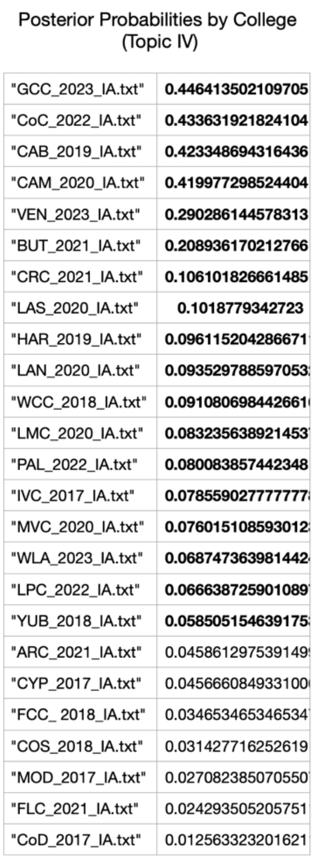
Posterior Probabilities by College
(Topic II)

"MOD_2017_IA.txt"	**0.57396449704142**
"YUB_2018_IA.txt"	**0.464948453608247**
"WLA_2023_IA.txt"	**0.403205398566006**
"WCC_2018_IA.txt"	**0.244926852288815**
"LAN_2020_IA.txt"	**0.111467008327995**
"HAR_2019_IA.txt"	**0.107836570663094**
"IVC_2017_IA.txt"	**0.106770833333333**
"PAL_2022_IA.txt"	**0.10188679245283**
"BUT_2021_IA.txt"	**0.09829787234042**
"FCC_ 2018_IA.txt"	**0.09488448844884**
"LAS_2020_IA.txt"	**0.08309859154929**
"CAB_2019_IA.txt"	**0.07987711213517**
"CYP_2017_IA.txt"	**0.07911576497963**
"VEN_2023_IA.txt"	**0.05647590361445**
"LMC_2020_IA.txt"	**0.05568581477139**
"CoC_2022_IA.txt"	0.04356677524429
"CRC_2021_IA.txt"	0.038087835211815
"ARC_2021_IA.txt"	0.03803131991051
"LPC_2022_IA.txt"	0.03688181056160
"MVC_2020_IA.txt"	0.03541076487252
"CAM_2020_IA.txt"	0.02648505486189
"GCC_2023_IA.txt"	0.02109704641350
"COS_2018_IA.txt"	0.01945525291828
"FLC_2021_IA.txt"	0.01239464551313
"CoD_2017_IA.txt"	0.01094224924012



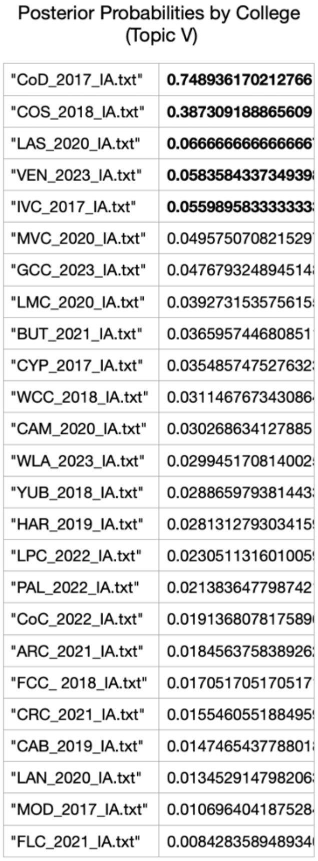
Posterior Probabilities by College
(Topic III)

"FLC_2021_IA.txt"	**0.612791274169559**
"CYP_2017_IA.txt"	**0.455206515415939**
"FCC_ 2018_IA.txt"	**0.435643564356436**
"ARC_2021_IA.txt"	**0.29586129753915**
"HAR_2019_IA.txt"	**0.264902880107167**
"CRC_2021_IA.txt"	**0.227361057131753**
"MVC_2020_IA.txt"	**0.201605288007554**
"IVC_2017_IA.txt"	**0.110677083333333**
"PAL_2022_IA.txt"	**0.108176100628931**
"LAN_2020_IA.txt"	**0.08584240871236**
"LPC_2022_IA.txt"	**0.07753562447611**
"VEN_2023_IA.txt"	**0.06212349397590**
"LAS_2020_IA.txt"	0.04366197183098
"WLA_2023_IA.txt"	0.03838043019822
"CoC_2022_IA.txt"	0.03460912052117
"BUT_2021_IA.txt"	0.03361702127659
"LMC_2020_IA.txt"	0.033411488862837
"WCC_2018_IA.txt"	0.03303445021236
"CAB_2019_IA.txt"	0.032258064516129
"MOD_2017_IA.txt"	0.02412380518889
"CAM_2020_IA.txt"	0.01891789632992
"YUB_2018_IA.txt"	0.01649484536082
"GCC_2023_IA.txt"	0.01645569620253
"COS_2018_IA.txt"	0.01586351391798
"CoD_2017_IA.txt"	0.01560283687943



Posterior Probabilities by College
(Topic IV)

"GCC_2023_IA.txt"	**0.446413502109705**
"CoC_2022_IA.txt"	**0.433631921824104**
"CAB_2019_IA.txt"	**0.423348694316436**
"CAM_2020_IA.txt"	**0.419977298524404**
"VEN_2023_IA.txt"	**0.290286144578313**
"BUT_2021_IA.txt"	**0.208936170212766**
"CRC_2021_IA.txt"	**0.106101826661485**
"LAS_2020_IA.txt"	**0.1018779342723**
"HAR_2019_IA.txt"	**0.09611520428667**
"LAN_2020_IA.txt"	**0.09352978859705**
"WCC_2018_IA.txt"	**0.09108069844266**
"LMC_2020_IA.txt"	**0.08323563892145**
"PAL_2022_IA.txt"	**0.080083857442348**
"IVC_2017_IA.txt"	**0.07855902777777**
"MVC_2020_IA.txt"	**0.07601510859301**
"WLA_2023_IA.txt"	**0.06874736398144**
"LPC_2022_IA.txt"	**0.06663872590108**
"YUB_2018_IA.txt"	**0.05850515463917**
"ARC_2021_IA.txt"	0.04586129753914
"CYP_2017_IA.txt"	0.04566608493310
"FCC_ 2018_IA.txt"	0.03465346534653
"COS_2018_IA.txt"	0.031427716252619
"MOD_2017_IA.txt"	0.02708238507055
"FLC_2021_IA.txt"	0.024293505205751
"CoD_2017_IA.txt"	0.012563323201621



Posterior Probabilities by College
(Topic V)

"CoD_2017_IA.txt"	**0.748936170212766**
"COS_2018_IA.txt"	**0.387309188865609**
"LAS_2020_IA.txt"	**0.066666666666666**
"VEN_2023_IA.txt"	**0.058358433734939**
"IVC_2017_IA.txt"	**0.055989583333333**
"MVC_2020_IA.txt"	0.049575070821529
"GCC_2023_IA.txt"	0.047679324894514
"LMC_2020_IA.txt"	0.039273153575615
"BUT_2021_IA.txt"	0.036595744680851
"CYP_2017_IA.txt"	0.035485747527632
"WCC_2018_IA.txt"	0.031146767343086
"CAM_2020_IA.txt"	0.030268634127885
"WLA_2023_IA.txt"	0.029945170814002
"YUB_2018_IA.txt"	0.028865979381443
"HAR_2019_IA.txt"	0.028131279303415
"LPC_2022_IA.txt"	0.023051131601005
"PAL_2022_IA.txt"	0.021383647798742
"CoC_2022_IA.txt"	0.019136807817589
"ARC_2021_IA.txt"	0.018456375838926
"FCC_ 2018_IA.txt"	0.017051705170517
"CRC_2021_IA.txt"	0.015546055188495
"CAB_2019_IA.txt"	0.014746543778801
"LAN_2020_IA.txt"	0.013452914798206
"MOD_2017_IA.txt"	0.010696404187528
"FLC_2021_IA.txt"	0.008428358948934

APPENDIX C: POSTERIOR PROBABILITIES BY TERM AND TOPIC

Figure 19 *Term Posterior Probabilities for Student-Centered Educational Process*

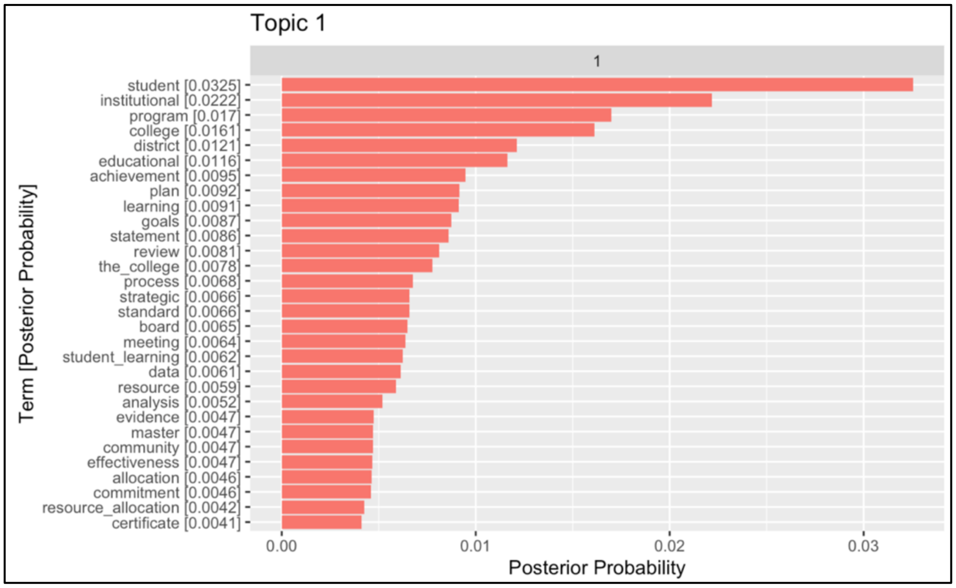


Figure 20

Term Posterior Probabilities for Deliberate Degree Delivery Process

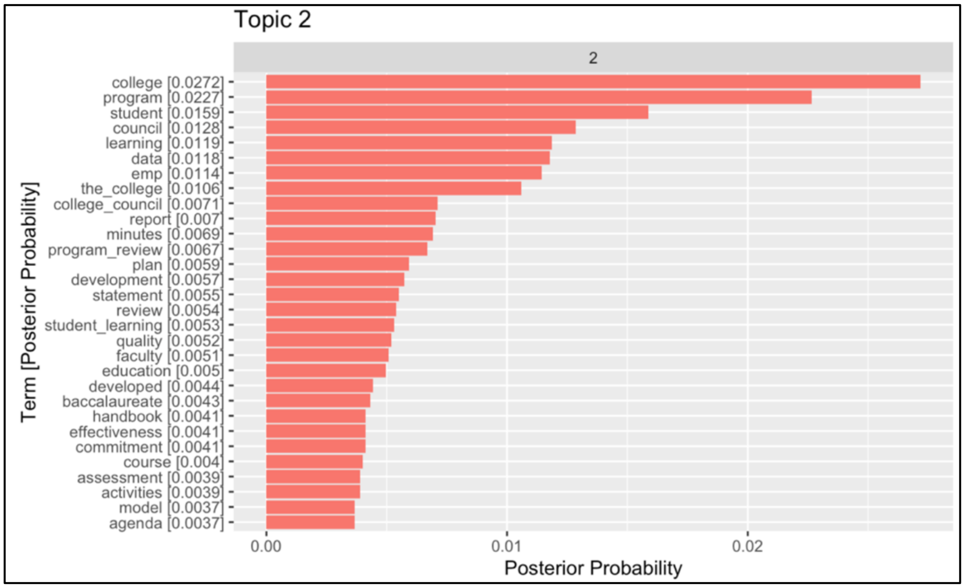
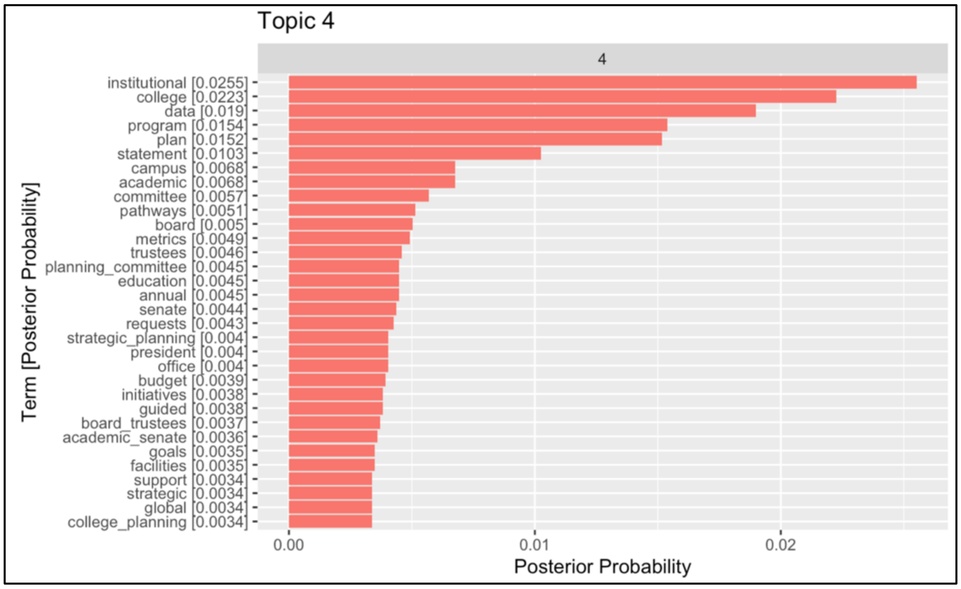


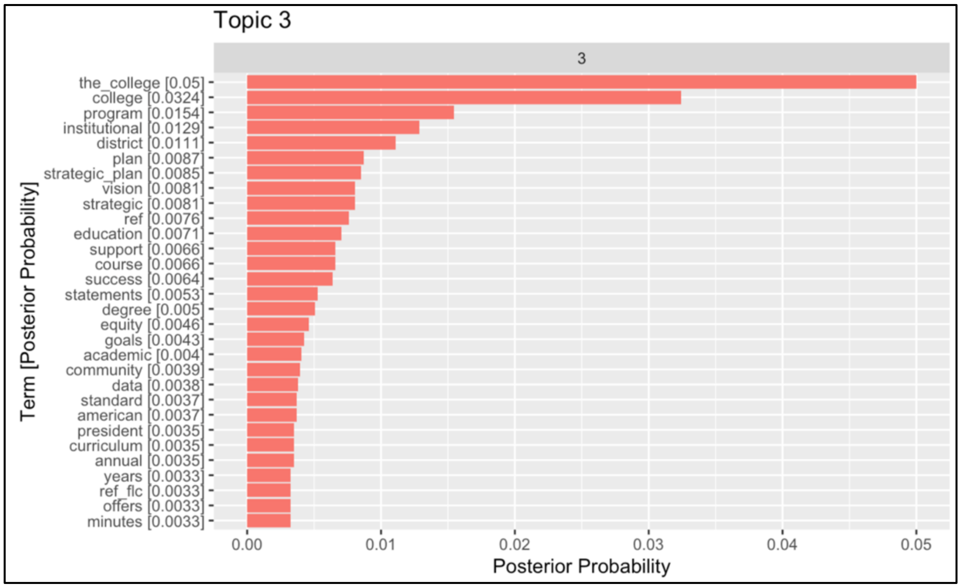
Figure 21

Term Posterior Probabilities for Stakeholder Coordination of Strategic Planning



Note: The topic numbers for topics 3 and 4 were reversed when the table was created. The terms and posterior probabilities are correct and match the topic numbers shown in LDAvis.

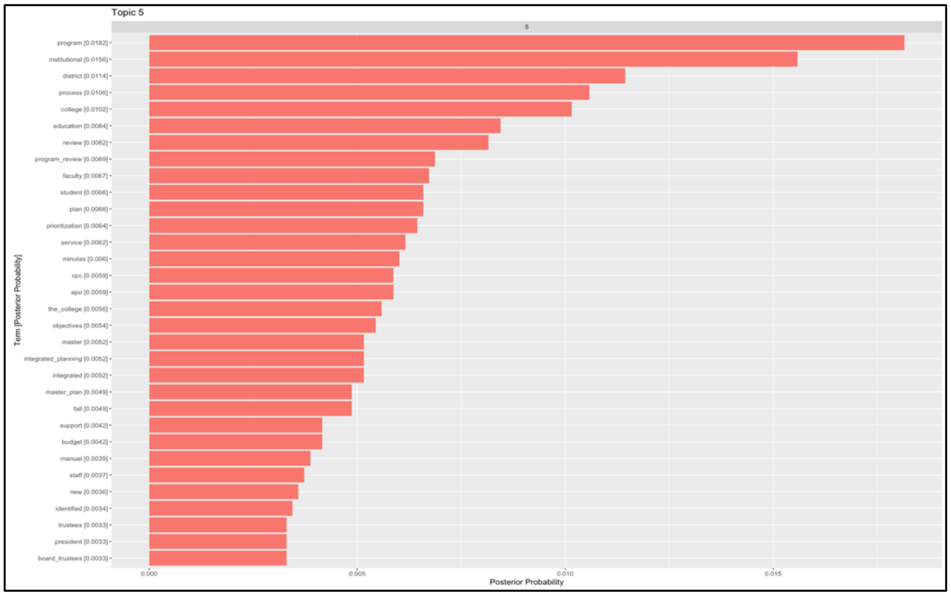
Figure 22 *Term Posterior Probabilities for Strategic Planning for Institutional Goals*



Note: The topic numbers for topics 3 and 4 were reversed when the table was created. The terms and posterior probabilities are correct and match the topic numbers shown in LDAvis.

Figure 23

Term Posterior Probabilities for Institutional Planning and Assessment


Topic 5
program [0.0182]
institutional [0.0156]
district [0.0114]
process [0.0106]
college [0.0102]
education [0.0084]
review [0.0082]
program_review [0.0069]
faculty [0.0067]
student [0.0066]
plan [0.0066]
prioritization [0.0064]
service [0.0062]
minutes [0.006]
cpc [0.0059]
apo [0.0059]
the_college [0.0056]
objectives [0.0054]
master [0.0052]
integrated_planning [0.0052]
integrated [0.0052]
master_plan [0.0049]
fall [0.0049]
support [0.0042]
budget [0.0042]
manual [0.0039]
staff [0.0037]
new [0.0036]
identified [0.0034]
trustees [0.0033]
president [0.0033]
board_trustees [0.0033]
Term [Posterior Probability]
Posterior Probability