AFRL-RH-WP-TR-2019-0100

Literature Review of Computer-Assisted Text Analysis Research, Software, Analytical Techniques, and Best Practices

Michael A. Campion, Ph.D.
Emily D. Campion, Ph.D.

Campion Consulting Services
403 West State St.
West Lafayette, IN 47907-2056

October 2019
Interim Report

DISTRIBUTION A: Approved for public release; distribution unlimited.

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING
AIRMAN SYSTEMS DIRECTORATE
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

AFRL-RH-WP-TR-2019-0100 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

CARRETTA.THOMAS.R.1228984929 Digitally signed by CARRETTA.THOMAS.R.1228984929 Date: 2020.03.05 07:51:14 -05'00'

THOMAS R. CARRETTA
Work Unit Manager
Collaborative Interfaces and Teaming Branch
Warfighter Interface Division
711th Human Performance Wing
Air Force Research Laboratory

WEBB.TIMOTHY.S.1085032239 Digitally signed by WEBB.TIMOTHY.S.1085032239 Date: 2020.03.17 15:38:55 -04'00'

TIMOTHY S. WEBB
Chief, Collaborative Interfaces and Teaming Branch
Warfighter Interface Division
711th Human Performance Wing
Air Force Research Laboratory

CARTER.LOUISE.ANN.1230249128 Digitally signed by CARTER.LOUISE.ANN.1230249128 Date: 2020.03.20 07:02:47 -04'00'

LOUISE A. CARTER, Ph.D., DR-IV
Chief, Warfighter Interface Division
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE ***(DD-MM-YY)*** 4-10-19	**2. REPORT TYPE** Interim	**3. DATES COVERED *(From - To)*** 3-14-19 to 7-15-19

4. TITLE AND SUBTITLE
Literature Review of Computer-Assisted Text Analysis Research, Software, Analytical Techniques, and Best Practices

5a. CONTRACT NUMBER
FA8650-18-F-6828

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)
Michael A. Campion, Ph.D.
Emily D. Campion, Ph.D.

5d. PROJECT NUMBER
5329

5e. TASK NUMBER

5f. WORK UNIT NUMBER
H0SA

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Campion Consulting Services
403 West State St.
West Lafayette, IN 47907-2056

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory
711th Human Performance Wing
Airman Systems Directorate
Warfighter Interfaces Division
Collaborative Interfaces and Training Branch
Wright-Patterson AFB, OH 45433

Infoscitex Corporation
4027 Colonel Glenn Highway
Suite 210
Dayton, OH 45431

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
711 HPW/RHCC

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RH-WP-TR-2019-0100

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A. Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES
Subcontract number: FPH02-S031/190789 88ABW-2019-5796, Cleared 05 December 2019

14. ABSTRACT

The purpose of this study was to conduct a comprehensive literature review to summarize the state of the research on the use of computer-assisted text analysis (CATA) to measure knowledge, skills, abilities, and other characteristics (KSAOs) for applications in hiring, recruiting, evaluation, and training. A comprehensive review of the CATA research in management journals identified 242 articles. Their findings were summarized in terms of types of studies, textual material, analyses, construct validity methods, and software. A series of questions was addressed regarding the relevance of this method to the Air Force including: What attributes have been measured? What research has been conducted in the context of employment? What evidence supports effectiveness? What can be learned from other disciplines that use similar techniques? What are the main challenges in the Air Force's intended use? What are the recommendations for using CATA by the Air Force?

15. SUBJECT TERMS
Computer-aided text analysis; Computer-assisted text analysis; CATA; text analysis; machine learning; construct validity; content analysis; sentiment analysis, hiring, recruiting, evaluation, training

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT:**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** (Monitor)
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	SAR	92	Thomas R. Carretta **19b. TELEPHONE NUMBER** *(Include Area Code)* (937) 713-7143

TABLE OF CONTENTS

SECTION PAGE

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

The purpose of this document is to conduct a comprehensive literature review summarizing the state of the research on the use of computer-assisted text analysis (CATA) to measure knowledge, skills, abilities, and other characteristics (KSAO). Information from this review is intended to inform the building of tools and methods for staffing, performance evaluation, and training outcome measurement. This includes research topics, methods, analytical techniques, findings, software, and best practices. The emphasis is on what can be learned about construct validity (understanding the meaning of the constructs measured) because advanced analytic techniques are often criticized as a "black box." The review is divided into three main topics.

1. Description of Existing Literature

The methodology followed the customary steps for conducting a comprehensive review of the research literature. A total of 242 relevant articles were identified and included. The review revealed that there are three bodies of literature on CATA. First, the management literature is the primary focus because it tends to emphasize construct measurement. Second, there is the literature using text mining in other disciplines, but the lack of focus on construct validity makes this literature less useful to the project (the so-called "black box" problem) and so it is only reviewed generally. Third, there is a large literature using CATA to measure writing skill in education, which will only be relevant if the project requires measurement of writing skill.

This report summarizes the findings, but an Excel spreadsheet titled, Literature Review Table Supplemental Information, is attached to this technical report in Appendix B (Campion & Campion, 2019); the article citations are in Appendix A, and a list of videos, websites, and podcasts on the software are in Appendix B. As a backdrop to our review, the following section of the report briefly summarizes the various ways CATA can be conducted.

2. Summary of Findings

We summarize the type of studies in the literature in terms of frequencies, and then we provide a narrative review of the nature and findings for each type.

Types of Studies: Three types of studies emerged in the literature review: (1) studies using qualitative CATA methods only (28.1%), which usually used inductive, grounded theory approaches; (2) studies that used both qualitative and quantitative methods (62.4%), with sentiment analysis often yielding the quantitative component; and (3) reviews of the literature (9.5%) that introduced text mining, provide recommendations on how to conduct CATA, presented specific techniques for conducting CATA, or summarized the use of CATA to measure constructs.

Types of Textual Data: CATA has been used to analyze virtually all types of textual data (e.g., reports, transcripts of interviews and phone calls, news articles, press releases, open-ended responses on surveys, archival data, organizational documents, observations, online reviews, messages, tweets, etc.). In more than half of studies reviewed, pre-existing text data were examined (i.e., was not collected specifically for the study). This fact suggests that CATA may be able to identify new constructs because it taps into a new data source, and some of the problems with purposefully collected data (like impression management) may be avoided. However, if text data are collected to measure a particular attribute, the questions should ask about that attribute as opposed to inferring it from other indirect information. For example, if the goal is to measure past leadership accomplishments, it is better to ask about past leadership accomplishments and text mine the responses than to text mine indirect text data such as personal statements. This is because the direct approach is more likely to solicit relevant and complete data on leadership, while the personal statement may focus on life goals and not discuss past accomplishments.

Types of Text Analyses: We developed a typology for organizing the approaches to conducting content analysis using CATA. The typology classifies the types of analyses in terms of two dimensions. One dimension is whether the analysis required low or high human intervention. The other dimension is whether the analysis uses a low or high degree of computer automation. This results in four cells of types of analyses: (1) dictionary-based analyses that require a low amount of human intervention because they are usually based on existing dictionaries and low automation because they rely largely on basic word searches and simple counts, which includes sentiment analyses; (2) categorization-based analyses that require a higher amount of human intervention to categorize content largely based on human judgment and little automation; (3) unsupervised machine learning that requires low human intervention but high automation because the computer identifies and learns the structure of the data using one or more techniques (e.g., natural language processing, latent semantic analysis) with little human intervention; and (4) supervised machine learning that requires high human intervention and high automation because the researcher guides the analysis. Together, these categories describe the range of CATA in the literature as well as the range of choices available to researchers. Historically, the literature has used mostly the simpler and less sophisticated approaches to CATA, but that is likely to change as researchers recognize the value of the more sophisticated approaches and the software becomes more available.

Types of Construct Validity Methods: We found it useful to distinguish among five types of validity evidence to best interpret this literature:

1. Evidence based on clearly defining the content domain. This is considered content-based evidence of validity by all testing authorities. All studies in the literature review included this type of evidence in one form or another. Thus we did not code it separately because it did not distinguish between studies.
2. Evidence based on using subject matter expert (SME) judgments to interpret categories. This is a central component to content validity, but studies in our review varied on whether and how they used this type of evidence to inform construct validity.
3. Evidence based on the internal structure of statistical measures. This type of evidence refers to statistical indicators of proper measurement and may include reliability or other psychometric statistical quality indicators, factor structure, or other information about the relationships among items or sections of the measure. Studies on CATA varied widely on whether this information was used to support construct validity and what types of statistical indicators were used.
4. Evidence based on relationships with other measures. Historically called "construct" validity, studies of CATA varied widely in their use of relationships with other measures.
5. Evidence based on predicting outcomes or consequences. This type of evidence is often called "criterion-related" validity. Again, studies of CATA varied widely in

their use of criteria to interpret construct validity, so coding of this topic may be informative to the Air Force purposes.

We discuss each type of method (except #1) and present examples of good practices of each from our review of the literature.

Types of Software: We started by identifying the types of software in the literature and counted the frequencies of their use. The most commonly used types of software for CATA were those developed for the key purpose of content analysis, or reducing large amounts of qualitative data into a digestible number of representative categories. Nvivo (60 studies, 24.6%) and Atlas.ti (18 studies, 7.4%) were popular. An equally popular software in our review was Linguistic Inquiry and Word Count (LIWC) (56 studies, 23%), which uses dictionaries mainly for sentiment analysis. DICTION (13 studies, 5.3%) is another software that uses built-in dictionaries to score text data, primarily for sentiment analysis. The more sophisticated software in the literature include R (13 studies, 5.3%), Python (Natural Language Toolkit (NLTK) (seven studies, 2.9%), and Statistical Package for Social Sciences (SPSS) Modeler (3 studies, 1.2%), which allow researchers to apply advanced analytics such as Latent Semantic Analysis (LSA) and Natural Language Processing (NLP). An important observation is that fewer studies than expected used the more sophisticated approaches incorporating LSA or NLP (52 studies, 23.9%) and only a few used commercial software products like SPSS (three studies) and Statistical Analysis System (SAS) (one study). We researched and evaluated these eight software packages in terms of each of the topics of relevance to this project: software name, source to acquire & cost, function, information on efficacy (e.g., validity), information on usability (e.g., difficulty to learn; technical support), likely applicability to the available Air Force data, likely applicability for measuring attributes relevant to the Air Force, anticipated advantages, anticipated disadvantages or problems, and overall evaluation of potential usefulness.

3. Answers to Questions of Importance to the Air Force

In this section, we provide answers to the questions of importance to the Air Force.

What attributes have been measured? We identified all the constructs in the literature, and then sorted them into categories of similar constructs. We found that nearly 200 attributes have been measured that could be summarized into seven categories of attributes, plus two additional related categories: sentiment (52 studies, 24%), cognition (51 studies, 23%), organizational characteristics that influence stakeholder psychological reactions (27 studies, 12%), behavior/skill (21 studies, 10%), personality/orientations (17 studies, 8%), language that influences human psychological reactions (17 studies, 8%), leadership (5 studies, 2%), processes that reflect relationships among attributes (5 studies, 2%), and unique attributes (17 studies, 8%).

What research has been conducted in the context of employment? We identified 29 articles that appeared to measure constructs potentially relevant to KSAOs for staffing decisions as well as other staffing issues (e.g., recruitment and turnover). They are described in terms of how they used CATA to measure KSAOs and how a similar approach might be useful in the Air Force context.

Of special note is that very limited research has been conducted on the topic of subgroup differences in text mining. Only four studies could be found that commented on subgroup differences, which we discuss. There were also several presentations at the 2019 Society for

Industrial and Organizational Psychology conference on machine learning that included discussion of subgroup differences and the potential for adverse impact, which we also summarize.

What evidence supports effectiveness? The evidence of effectiveness have been of two types: Evaluation and validation. We described the aforementioned types of validation evidence under types of construct validity. The total number of studies on CATA in peer-reviewed journals and the amount of research using each type of construct validity evidence suggests that there is substantial evidence of validity. Evaluation evidence includes a wide range of techniques focused on the quality of the computer model itself, rather than its usefulness. The primary types of evaluation evidence included the following: Cross-validation, accuracy of classification, psychometric indices, interpretability, and operational considerations.

What can be learned from other disciplines that use similar techniques? Potentially useful observations from that literature include the following: (1) The machine learning research outside management often includes both quantitative variables based on numeric data and textual variables based on text mining; (2) The focus in the other literatures on prediction of outcomes over construct validity reveals that there are many other statistical models that might improve prediction; (3) The range of outcomes predicted in those other literatures might give the Air Force ideas as to the broader applicability of these techniques; (4) Researchers in other fields will often make adjustments to the distributions of the data, including both trimming and imputing data, and using assumed distributions; (5) Other literatures show that the sources, styles, and amount of data are seemingly endless, some of which might be of value to the military.

What are the main challenges in the Air Force's intended uses? Some of the challenges discussed include: (1) Identifying viable applications: This will depend on many considerations, some of which are whether textual data are currently collected from the applicants, whether the textual data collected likely contain information on job-related constructs, whether there is enough variance in the data to be useful for selection, and whether a large enough sample of data can be collected to create the CATA model or whether Subject Matter Experts (SME) could write illustrative text used to train the model; (2) Deciding on software to use; (3) Learning the software; (4) Training the software; (5) Validating; (6) Updating the model to accommodate legal and social changes; (7) Working out operational details; and (8) Communicating to candidates.

What are the recommendations for using CATA by the Air Force? Many recommendations are made explicitly or implicitly throughout the report, both for using CATA and many other topics. This section summarizes our recommendations regarding the approach to use. We recommend:

1. In terms of a bottom-line, CATA should be used as an approach to measuring KSAOs for employment decisions.
2. Use more sophisticated methods than CATA methods in most of the management literature to date. Specifically, use approaches to text mining that evaluate strings of words and relationships among words like LSA and NLP, as opposed to the more common simple single word and phrase-based approaches.
3. Use approaches that allow training of the CATA model.
4. Do not ignore the use of word dictionaries because they offer potential advantages.

5. Consider the various ways to strategically select or create corpuses (corpora) for developing CATA models.
6. Consider sentiment analysis as it might be appropriate.

1.0 INTRODUCTION

The purpose of this report is to conduct a comprehensive literature review to summarize the state of the research on the use of computer-assisted text analysis (CATA) to measure knowledge, skills, abilities, and other characteristics (KSAO) (e.g., writing skills, level of motivation for a given job or job type, personality, temperament, behavioral intentions). In other words, the primary purpose of the literature review is to determine what constructs have been measured and what methodologies have been used that are relevant to the goals of the project. The goals of the project are primarily to use text analysis to measure job-related KSAOs to build staffing tools (e.g., analyze statements of skills or other application information), but will also include some related accession uses (e.g., summarizing survey data on reasons for turnover) and possibly other uses such as improving performance evaluation and training outcome measurement. This will be accomplished in part by conducting a comprehensive literature review of CATA research, software, analytical techniques, and best practices.

We will use the terms text analysis, text mining, and CATA interchangeably throughout the report. When we are referring to a specific software or analysis technique, we will specify it by its name or other descriptive terms. The current report describes the methodology, findings, and recommendations. Accompanying documents contain the comprehensive table summarizing each article or other document included in the review, as well as electronic copies. We encourage the reader to utilize these resources to gather further information and gain a deeper understanding of the source literature.

The scope of the review is meant to be all-inclusive of any literature relevant to the topic. This scope was maintained throughout the project, but the relevance of the various bodies of literature narrowed for three reasons. First, because this is such a relatively new and rapidly evolving area of research, most of the relevant literature has been published recently (from 2016 to 2019). Second, there is a vast literature on machine learning that did not include CATA, and thus was not directly relevant to the purpose of the review. Third, the goals of the project mean that the focus will be on the literature that bears on construct measurement of human attributes, which is primarily the research in management (psychology, organizational behavior, and human resources) as opposed to other disciplines. The review will also scan the text mining literature in other disciplines for lessons that might be relevant to the project goals, such as methodologies, but not review them comprehensively. These and other refinements to the scope will be described in more detail below.

1.1 Plan for the Review

The review is divided into three main topics: (1) A description of the existing literature, including the methodology followed, the bodies of literature discovered, and the information collected; (2) a summary of the findings organized around the various types of distinctions between the articles, with a separate section for each, and (3) answers to the questions of importance to the Air Force, which include the insights discovered and the recommendations for using CATA.

2.0 METHOD, ASSUMPTIONS, AND PROCEDURES

The methodology followed the customary steps for conducting a comprehensive review of the research literature. Steps taken were as follows.

1. We identified and used all electronic databases of potential relevance to the topic (e.g., Academic Search Complete, Applied Science & Technology Source, Business Source Complete, Communication & Mass Media, Education Source, Education Resources Information Center (ERIC), Google Scholar, Military & Government Collection, PsychArticles, and PsychINFO). In addition, we identified 11 academic, peer-reviewed journals dedicated to data science that were searched, including *Artificial Intelligence*, *Big Data*, *Big Data Research*, *Computational Statistics & Data Analysis*, *Foundations and Trends in Machine Learning*, *International Journal of Business Intelligence and Data Mining*, *International Journal of Data Science and Analytics*, *Journal of Big Data*, *Machine Learning*, *SIGKDD Explorations*, and *Data Science Journal*.
2. We identified and used all keywords of potential relevance to the topic (e.g., automated essay scoring, automated writing evaluation, computer-aided text analysis, computer-assisted text analysis, electronic essay scoring, latent semantic analysis, machine learning, natural language processing, and text mining).
3. We reviewed the titles and abstracts of potentially relevant articles.
4. We cross-referenced and forward-searched relevant articles to identify additional relevant articles.
5. We read and summarized all relevant articles in a table that included all the potentially useful information from the articles based on the purposes of the review (see below).
6. We studied and analyzed the table and associated articles to extract the findings, conclusions, and recommendations relevant to the project as described in the current report.

Following these steps, the review occurred in two phases as depicted below. Phase 1 was the original inclusive database search, wherein we searched the databases in #1 above simultaneously, which identified nearly 15,000 articles. However, most were not relevant because they were on machine learning without a text analysis component. In other words, they only analyzed numeric and not textual data. Focusing only on text analysis articles reduced the number to 1,647. Reviewing the articles and benchmarking with other scholars at professional conferences and workshops on CATA revealed that only the most recent articles (from 2016 to 2019) would be relevant because of the newness and rapidly changing nature of this approach to research methodology. That reduced the number of articles to 936. Reviewing the abstracts of those articles revealed that only 108 were relevant. Reading those articles in detail reduced the final number of relevant articles to 84.

In Phase 2, we discovered from initially reviewing the literature that another area of relevant literature existed that was not identified by the original search. This was literature attempting to measure management-related constructs, which were usually human attributes and would be highly relevant to our purposes, but was identified in the literature by the terms “sentiment analysis” and “content analysis” rather than “text analysis” and the other keywords above. A recent review of that literature by Short, McKenny, and Reid (2018) contained 144 articles of which 102 were in management-related journals measuring management-related constructs. We updated that literature review for the years 2017-2019 using the same terms as Short et al. (e.g., computer-aided text analysis, CATA, computerized content analysis, LIWC and other software that conducts sentiment analysis, Nvivo and other software that conducts content analysis, and articles citing Short's review). This identified 56 additional articles in management-related journals and 21 not in management related journals and not measuring management-related constructs. These 158 articles (102 and 56) plus the 84 from the original search totaled the 242 articles included in the present literature review. Figures 1 and 2 below depict the search phases pictorially showing each step.

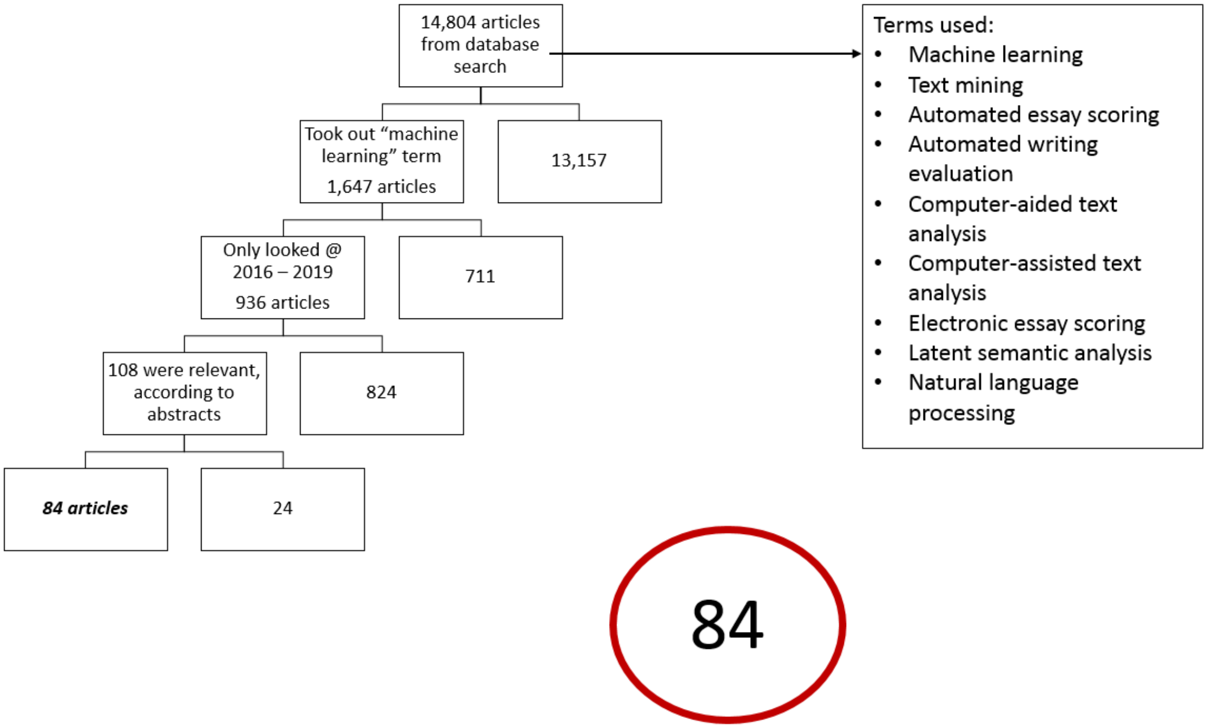
14,804 articles from database search
Took out "machine learning" term
1,647 articles
13,157
Only looked @ 2016 – 2019
936 articles
711
108 were relevant, according to abstracts
824
84 articles
24
84
Terms used:
Machine learning
Text mining
Automated essay scoring
Automated writing evaluation
Computer-aided text analysis
Computer-assisted text analysis
Electronic essay scoring
Latent semantic analysis
Natural language processing

Figure 1: Illustration of Phase 1 of Literature Search Using Databases

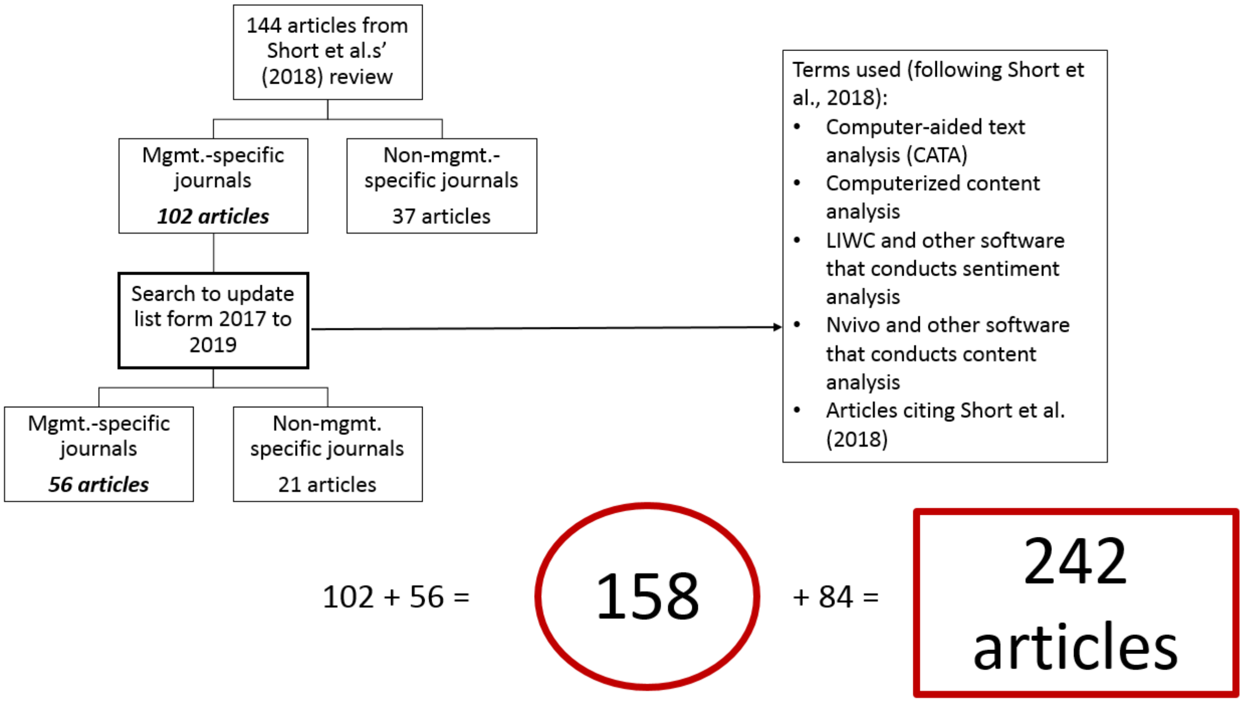
144 articles from Short et al.s' (2018) review
Mgmt.-specific journals
102 articles
Non-mgmt.-specific journals
37 articles
Search to update list form 2017 to 2019
Mgmt.-specific journals
56 articles
Non-mgmt. specific journals
21 articles
Terms used (following Short et al., 2018):
• Computer-aided text analysis (CATA)
• Computerized content analysis
• LIWC and other software that conducts sentiment analysis
• Nvivo and other software that conducts content analysis
• Articles citing Short et al. (2018)
102 + 56 =
158
+ 84 =
242 articles

Figure 2: Illustration of Phase 2 of the Literature Search in Managemment Journals for Construct Measurement Using Content and Sentiment Analyusis

2.1 Bodies of Literature

Broadly speaking, the review revealed that there are three bodies of literature on CATA. First, there is the relevant literature we described above. We will call it the "management literature" because it was published primarily in management-related journals, which tend to focus more on construct measurement. Although most of this literature is highly relevant to the core disciplines related to staffing (psychology, organizational behavior, and human resources), we included articles from journals in other areas of management (e.g., marketing, operations, finance, etc.) if they focused on measuring human constructs. Most of these articles only use CATA to perform sentiment analysis or content analysis, which are fairly basic approaches to text mining. Sentiment analysis seeks to measure affective reactions and other feelings, usually from word dictionaries. Content analysis seeks to summarize narrative data into categories of similar topics, usually based on common terms. However, a number of the articles use LSA and NLP, which involve extracting the meaningfulness of narrative information by analyzing the relationships among multiple words and is more helpful for construct measurement. Also, most of these do not focus on prediction, but some tried to predict other outcomes from the content categories identified.

Second, there is the literature using text mining that does not focus on construct validity in other disciplines, especially information technology and engineering, but also in other areas of management. There are many hundreds of articles here. They mainly try to predict outcomes from text data, but their lack of focus on construct validity makes them less useful to the project. In other words, they do not try to provide information on the meaning of the text. They only use the text to enhance the statistical prediction of the outcomes. For example, they might use past consumer comments to predict future consumer choices. Therefore, we reviewed them generally to pull out technical methodological lessons, but we did not analyze each article systematically like those in management-related journals.

Third, there is a large literature using CATA to measure writing skill in education. This was the first major area of literature using CATA and consists of many hundreds of articles. The focus on writing skill and not content or prediction makes this literature less useful to this project, unless the Air force decides to measure writing skill. In that case, we would rely on the many review articles in this literature. Considering the current focus of the project does not focus on writing skills, these articles will not be included in our analysis.

2.2 Information Collected

We captured the following list of variables for each article reviewed. Those bearing on construct validity are explained in more detail. Those, plus several other variables, will form the basis of our organization of the findings in the sections below.

A. Authors

B. Year

C. Journal

D. Title

E. Discipline

F. Type of Study

G. Abstract

H. Purpose of Study

I. Theory

J. Type of textual data used

K. Sample

L. Intended attributes measured

M. Type of software

N. Type of Analysis I (content or sentiment)

O. Type of Analysis II (categorization-based, dictionary-based, unsupervised machine learning, and supervised machine learning)

P. Brief summary of findings if content analysis used – such as number and description of categories extracted.

Q. Used SME judgments to interpret categories – whether used SME subjective judgments to interpret the constructs.

R. Calculated psychometrics on variables created – whether calculated reliabilities or other statistical quality indicators of the constructs measured.

S. Examined construct validity – whether correlated measures of the variables with other known measures to yield information on convergent and discriminant validity.

T. Predicted important outcomes – whether predicted any outcomes to bear on the importance and interpretation of the constructs measured.

U. Examined subgroup differences – whether subgroup differences were examined and a brief description of the findings.

V. Notes

W. Full citation

The hypotheses and findings are not coded because they will be reflected in the Abstract. The Excel table with this information is available at this link (Campion & Campion, 2019: XXX) and Appendix A contains the Literature Review List of Articles.

2.3 Basic Approaches to Text Analysis

Before describing the findings in the literature, it might be helpful to identify the various ways text mining can be conducted as a backdrop. We believe the range of approaches include, but are not limited to, the following:

1. The most basic approach to text analysis is traditional human judgment-based content analysis. Here, human judges such as SMEs read and categorize the sample of textual data based on similarity of the content. This technique is sometimes called a "Q Sort." After sorting, the human judges read the data sorted in each category and assign descriptive labels. Often, they will count the number of responses in each category as an indication of its importance. This is not computer assisted, other than perhaps to record the information. However, it provides an important perspective because human judges play a role in the more automated approaches below.

2. Using basic automation to help identify content by simply counting the frequencies of various words and helping categorize them. They may also help visually display the results such as in "word clouds." This approach can be purely empirical by only using the variables identified by the software, or the variables can be improved by modifying the categories by the researcher. This involves reviewing the concepts extracted and combining or separating them based on their meaning in the context of the study, just like traditional content analysis. The computer does not know that different terms may be synonyms, but most software will allow the researcher to tell that to the computer. The initial variables extracted can be modified using other analyses to reduce the data, such as factor or cluster analyses. SMEs might also be used to help impart meaning to the categories derived or to check the coding and make modifications. Simple scores can be derived for each category based on the number of data points it contains. This use of CATA has been very common in the management literature.

The most common software packages that conduct simple content analysis are Nvivo and Atlas.ti. Although they help suggest potential categories by identifying the common words, they serve as more of a data management tool for qualitative (text) data and involve little automation. They are word processing tools that facilitate text analysis by allowing the researcher to easily sort and keep track of the data, but provide little assistance in identifying the underlying meaning of the data compared with more automated methods described below.

3. Using rationally developed data dictionaries. This is most often used for sentiment analysis based on existing dictionaries, but researchers can also create their own dictionaries or use it for content analysis. This is much like a keyword search where the researcher identifies all the relevant terms and then searches the documents for these words. The difference is that the words have been identified as being reflective of various sentiments based on prior research. The sentiments can be as simple as positive/negative or highly complex and nuanced (e.g., specific attitudes and dispositions). The number of words identified and the frequency of their use usually provide the measure of the attribute. Studies using dictionaries are more deductive because researchers are either trying to measure particular attributes for which the dictionaries have been validated *a priori*, or they are developing their own dictionary based on a known construct that is also known *a priori*. Many dictionaries are available publicly to assess various sentiments or emotions (especially the LIWC and the DICTION software packages). Using CATA to identify sentiments has been a very common use in the management CATA literature.

4. Using more advanced text mining software to identify the constructs underlying combinations of words that exist in a corpus (e.g., set of documents) such as LSA and NLP techniques. They involve extracting the meaningfulness of narrative information by analyzing the relationships among multiple words, as described in more detail on the software review section of this report. Advanced techniques such as LSA and NLP can be used in addition to data dictionaries and basic automation (e.g., word clouds), and some software programs will combine all three. Common software programs that perform all these analyses include R, Python, SPSS, and SAS.

5. Using text mining software to identify concepts and combinations of words based on their predictiveness of some criterion (e.g., job performance, training performance, other outcomes, etc.). This approach starts with the approaches above, but then identifies the most useful concepts and variables from the large number extracted based on their empirical relationships with criteria. The number of variables extracted depends on the size and variation of the textual data in the corpus, but can range up to the hundreds or even thousands. This approach can be combined with training the computer as described above. This was the approach we used in our published article (Campion et al., 2016) and our ongoing work.

6. Using criterion data from SMEs to text mine against. Where criterion data do not exist, have SMEs score a set of written text samples and use their scores as the criterion (i.e., treat the SME ratings as the true scores). If well done, the sample of text scored would not have to be excessively large (e.g., in the hundreds) because the research protocol can be structured to ensure wide variance, reliability, and content validity. For example, in text mining accomplishment records from candidates for hiring as described in Campion et al. (2016), we used assessor scores as the criterion and we were able to achieve a correlation between the computer scored essays and the assessor scores that was as great as a single individual assessor interrater reliability (r = .60).

7. Using software that allows both numeric data variables as well as text mined variables to predict some outcome. For example, the predictive model (like a regression) might include quantitative candidate information, as well as text-mined narrative information collected from the candidate. The quantitative information can include years of work experience, types of degrees, test scores, and other application information. This is the approach we previously used

as cited above because including these other variables enhances prediction of outcomes (also see Sajjadiani, Sojourner, Kammeyer-Mueller, & Mykerezi, 2019).

8. Using software that scores writing skills as well as content. Although the major recent breakthrough in text mining is the scoring of content (e.g., job-related experience), the longest traditional use of text mining is in scoring writing skill, especially in educational settings. In many employment situations, using an automated writing skill assessment can be useful. However, because it does not focus on construct measurement beyond writing skill, it will not be the focus of the current review as noted previously. Nevertheless, our research into writing skill measurement software for other organizations has revealed several findings that may be useful to the Air Force. First, commercial software for scoring writing skills is not available for purchase. Second, three companies sell this as a service (ETS, Pearson, and Measurement Inc). Therefore, the Air Force could have samples of text scored for writing skill if desired, but could not have it as part of its hiring tools unless it contracted with these vendors. Third, our research has discovered that these scores show a strong correlation with essay length (r = .80), so they appear to measure verbal fluency that could be closely approximated by a simple measure of length. In other words, they appear to measure the skill in writing a lot of text quickly that is grammatically correct, as opposed to writing well on other dimensions (e.g., succinct, well structured, persuasive, interesting, or other elements of style). Fourth, because it is unknown whether available text mining software can measure writing skill, we recently did a study for another client. We used the SPSS software, which presumably scores content, to determine whether it could predict writing skill based on ratings by human graders in a large sample. We achieved a correlation of .48 using a completely untrained model. An untrained model is the initial model extracted by the computer. No humans have identified synonyms, combined similar categories, eliminated non-meaningful categories, or engaged in other forms of training. This suggests we might be able to score writing skill to some extent with our current software if that is of interest to the Air Force. However, it is likely that the SPSS software is picking up on writing skill, length, and content. Supervising the training of the model would allow us to focus it on one or the other. We are not aware of published studies using R or Python to measure writing skill.

With the exception of approach eight, which is out of scope, all of these approaches have been used in the research literature summarized below. We expect all of these approaches will be useful to the Air Force.

3.0 RESULTS AND DISCUSSION

Table 1 shows the frequency and percentage of each of the various types of distinctions between the articles in the review. The Excel table with the information on each article, titled Literature Review Table Supplemental Information, is attached to this technical report in Appendix B (Campion & Campion, 2019) and the list of articles is in Appendix A. The sections below describe and summarize the key findings within each type, with special emphasis on the goals of the Air Force to use CATA for personnel selection and related accession needs. As explained above, the total number of articles reviewed was 242. However, 23 were review articles or articles that presented information on how to use CATA in organizational research, so the percentages of other types are only based on the non-review articles (n = 219). The numbers will total more than 219 and the percentages will total more than 100% because some studies used several types.

Table 1: Frequencies of Types of Articles in the Literature Review

Type of Study	**Number**	**Percentage**
Qualitative and quantitative	151	62.0%
Qualitative only	68	28.0%
Review	23	10.0%
TOTAL	242	100.0%
Type of Textual Data	Number	Percentage
Reports	30	11.3%
Transcripts (interviews, phone calls)	81	30.4%
News-related documents (news articles, press releases)	31	11.7%
Open-ended responses	20	7.5%
Archival data and organizational documents	23	8.7%
Observations	19	7.1%
Online reviews, messages, or tweets	42	15.8%
Other (academic journal articles/abstracts, job postings, patient records)	20	7.5%
TOTAL	266	100.0%
Types of Text Analyses	Number	Percentage
Content versus Sentiment:		
Content analysis	146	66.7%
Sentiment analysis	55	25.1%
Content & Sentiment analyses	18	8.2%
TOTAL	219	100%
Degree of Automation:		
Dictionary-based analysis	109	49.8%
Categorization-based analysis	96	43.8%
Supervised machine learning	12	5.4%
Unsupervised machine learning	1	.5%
Supervised & Unsupervised machine learning	1	.5%
TOTAL	219	100.0%

Types of Construct Validity Methods	Number	Percentage
Predicted important outcomes	102	34.8%
Used SME judgments to interpret categories	93	31.7%
Calculated psychometrics on variables created	58	19.8%
Convergent and discriminant validation	40	13.7%
TOTAL	293	100.0%

Types of Software	Number	Percentage
Nvivo	60	24.8%
Linguistic Inquiry Word Count (LIWC)	56	23.1%
Did not report	23	9.5%
Atlas.ti	18	7.4%
DICTION	13	5.4%
R	13	5.4%
SentiWordNet	5	2.1%
MonoConc Pro 2.0	4	1.7%
Python	4	1.7%
Natural Language ToolKit (NLTK)	3	1.2%
AFINN sentiment lexicons	2	0.8%
Apache Solr	2	0.8%
Automap	2	0.8%
General Inquirer (GI)	2	0.8%
SPSS Modeler	2	0.8%
Textual Analysis Computing Tools (TACT)	2	0.8%
Textpak 4	2	0.8%
Valence Aware Dictionary and sEntiment Reasoner (VADER)	2	0.8%
VBPro	2	0.8%
WordNet	2	0.8%
Wordstat	2	0.8%
Only Used Once (Not Listed Individually Here)	21	8.7%
TOTAL	242	100.0%

3.1 Types of Studies

Three types of studies emerged in the literature review: (1) Studies using qualitative CATA methods only, (2) studies using both qualitative and quantitative methods, and (3) reviews of the literature. The distinction between the first and second types of studies is whether the researchers converted the qualitative data into quantitative data. That is, if the data are converted into more than just counts and categories, such as indices and metrics, and especially if they are used for statistical analysis like correlations, then it becomes both qualitative and quantitative. Incorporating other quantitative data also makes it both. Of the 242 articles, 68 (28.0%) were qualitative only and generally took inductive approaches using interviews with organizational informants (e.g., respondents, experts) to develop constructs and flesh out processes among constructs in management theory. Coding in qualitative studies can be done in a number of ways, but the primary approach is called *grounded theory*. According to Gephart (2004, p. 459):

> “*Grounded theorizing* (Glaser & Strauss, 1967) is the process of iteratively and inductively constructing theory from observations using a process of theoretical sampling in which emergent insights direct selection and inclusion of the “next” informant or slice of data. Grounded theory involves constant comparative analysis whereby groups are compared on the basis of theoretical similarities and differences.”

Scholars who use qualitative research often apply grounded theory with the intent to generate or refine theory (Edmondson & McManus, 2007). They begin with an inductive approach to gather information about the phenomenon and then build theory from it, sometimes deductively testing propositions as they are developed, or iterating between emerging theoretical ideas and the data in an inductive-deductive cycle to refine theory (e.g., Strauss & Corbin, 1990). This is in contrast to traditional deductive research that usually starts with an existing theory and then develops and tests hypotheses derived from that theory to determine the support.

This theoretical framework requires researchers to code as close to the data as possible. Recognizing that several ideas can be communicated in one spoken (and transcribed) sentence, researchers using this method first engage in open coding, or a coding ritual called “in vivo” based directly on words used by participants and are as similar to the terms used in the data as possible without applying external theoretical framings. Such line-by-line coding allows for the researchers to examine individual words and phrases (e.g., Locke, 2001). This is much like what was observed when extracting concepts in the example content analyses conducted for the Air Force as part of the current project. Next, in an analogous process to creating superordinate categories in traditional content analyses, researchers begin to combine the first-order codes into second-order codes (axial codes; Pratt, 2009) while iterating between the literature and the data in the first-order codes to develop the higher-order codes. This was also illustrated by creating categories of the concepts in the Air Force examples, although combinations were based on content similarity in the data rather than categories based on some prior theory. The final step is creating aggregate dimensions that are then used in the development of a model of the target topic. Caza, Moss, and Vough (Figure 1, 2018) illustrate the process of creating first-order codes that aggregate to second-order codes that aggregate to dimensions. This approach is generally used in studies that do not attempt to quantify the textual data (e.g., create ratings, scores, or word frequencies) but try to simply summarize the content and/or generate theory.

To illustrate a grounded theory type of study, consider a study where scholars examined the process through which U.S. Navy couples manage and navigate their work and home demands. Beckman and Stanko (2019) used this sample to expand boundary theory, which essentially states that individuals have boundaries or “mental fences” around their roles (e.g., employee, parent, spouse) and how individuals conceptualize the boundary among their roles and how they transition between them affects their performance in each role (Ashforth, Kreiner, & Fugate, 2000). Through in-depth interviews with 29 U.S. Navy couples, Beckman and Stanko coded qualitative data and developed a model to extend boundary theory with relational boundary work, which is what each individual engages in to simultaneously remain committed to the Navy while building resilience as a couple.

Of the 242 articles, 151 (62.0%) used both qualitative and quantitative analyses. Typically, this involved the use of CATA to quantify qualitative data. A common example was creating a sentiment estimate based on a word count of the number of words in the text that represented a certain sentiment (e.g., positive or negative). These estimates then were used to correlate with or predict important outcomes. For example, Love, Lim, and Bednar (2017) analyzed the sentiment of 200 news articles on Chief Executive Officers (CEO) and used this variable ("CEO media tenor") to predict firm reputation based on Fortune Magazine's "Most Admired Companies" surveys. They took a number of steps to create a quantitative measure of tenor. First, using LIWC's dictionaries of positive and negative words, they counted the number of positive and negative words in each response. Then, they created ratios of positive to negative words used and negative to positive words used. Finally, they coded articles as positive if the positive ratio was .65 or greater and negative if the negative ratio was .65 or greater. However, this is only one way of creating a sentiment score. In a similar study on media favorability of organizations, Bednar (2012) also used LIWC to evaluate news articles, but calculated the mean frequency of positive and negative words "from all articles about a sample firm in a given year" (p. 138). Bednar tested whether media favorability was affected by formal board independence and also whether media favorability affected CEO pay and likelihood of CEO dismissal.

Another way to measure sentiment is a simple percentage of the number of sentiment words (positive or negative) to the total number of words in a text. Wilson, DeRue, Matta, Howe, and Conlon (2016) applied this method in a study of emotional displays in negotiations. The positive emotional displays occurring in a negotiation were operationalized as the percentage of positive emotion words identified by LIWC (e.g., agree, enjoy, great, nice, perfect, thanks) within each negotiation transcript" (p. 1411). Wilson et al. then tested whether personality similarity between negotiators enhanced positive emotional display and whether positive emotional displays quickened agreement time and reduced perceptions of relationship conflict. They found support for these hypothesized relationships.

Finally, of the 242 articles, 23 (10.0%) were reviews. The reviews had somewhat different purposes, the most common of which were: (1) introducing text mining (e.g., Kobayashi, Mol, Berkers, Kismihok, & Hartog, 2018a, 2018b; Luciano, Mathieu, Park, & Tannenbaum, 2018), (2) providing recommendations on how to conduct CATA (e.g., Banks, Woznyj, Wesslen, & Ross, 2018; Grimmer & Stewart, 2013; McKenny, Short, & Payne, 2012; Medhat, Hassan, & Korashy, 2014), (3) presenting specific techniques that included CATA (e.g., Blei, Ng, & Jordan, 2003; Crayne & Hunter, 2018; Hannigan et al., 2019; Janasik, Honkela, & Bruun, 2009; Shortt & Warren, 2018; Slutskaya, Game, & Simpson, 2018), and (4) using CATA to measure constructs in organizational behavior and psychology research (e.g., Short, McKenny, & Reid, 2018), which was particularly useful to the present review because it identified many relevant articles.

In summary, there were three types of studies in the literature on CATA. More than a fourth have been grounded theory studies, which used purely qualitative CATA data to develop new theories. Almost two thirds have included both qualitative and quantitative data, usually by the addition of sentiment dictionary CATA data as the quantitative data to predict other outcomes. Finally, almost 10% of the studies are reviews promoting CATA or summarizing the CATA literature.

3.2 Types of Textual Data

Perhaps what makes CATA so promising for researchers and practitioners alike is that any and all written text (including transcripts or oral data) can be analyzed. The textual data used in the research literature included various types collected intentionally for the study (such as from interviews and open-ended responses on surveys) and types of existing data created for other purposes (such as letters to shareholders, news articles and press releases, and online reviews and social media). In terms of studies that collected data specifically for the research, the most common source was transcripts of interviews, focus groups, phone calls, conference calls, etc. (81 studies; 30.4%). Illustrative attributes from these types of text data include perceptions of fit (Chuang, Hsu, Wang, & Judge, 2015), humble leadership (Owens & Hekman, 2012), cognitions and emotions (Zuzul, 2019), and identity (Creed, DeJordy, & Lok, 2010; Giois, Price, Hamilton, & Thomas, 2010). Other less commonly used textual sources specifically collected for the research are open-ended responses (20; 7.5%) and observations (e.g., researcher recordings of behavior based on visual observations) (19; 7.1%). Illustrative attributes from these types of text data include team cognitive maps (Carley, 1997), boundary management tactics (Kreiner, Hollensbe, & Sheep, 2009), and positive-negative sentiment (Liang et al., 2016).

In terms of using existing data, the most commonly used textual source was online reviews, messages, and related online interactions (e.g., Twitter "tweets") (42; 15.8%) followed by reports (30; 11.3%), such as letters to shareholders and analyst reports. Examples of attributes measured by these sources were CEO characteristics such as narcissism (Buyl, Boone, & Wade, 2019) and entrepreneurial orientation (Engelen, Neumann, & Schmidt, 2016) or firm-specific attributes such as organizational culture (Pandey & Pandey, 2017) and organizational values (Kabanoff & Holt, 1996; Kabanoff, Waldersee, & Cohen, 1997). The next most common existing sources were archival data (e.g., resumes) and organizational documents (e.g., mission and value statements) (23; 8.7%). Researchers tended to use archival data or other organizational documents to measure attributes such as organizational change emergence (Wiedner, Barrett, & Oborn, 2017) and online reviews and messages to measure sentiment (Barlow, Berhaal, & Hoskins, 2018). Many researchers also took advantage of existing news-related documents including news articles and press releases (31; 11.7%), which can include important reputational information regarding CEOs and other organizational attributes (Gangloff, Connelly, & Shook, 2016; Quigley, Hubbard, Ward, & Graffin, 2019). Twenty articles (7.5%) used other types of existing textual data that did not fall within these categories such as abstracts of academic journals, job postings, and patient records.

CATA can be used to better analyze data collected as part of traditional quantitative data collection (like open-ended questions in surveys), but it is noteworthy that most of the studies applying CATA used existing text data not originally collected for the purpose of text analysis research (cumulatively 55.0%). This means CATA is being applied to entirely new, previously untapped data sources. That is, it is being used often to analyze new types of data, not just to better analyze existing data commonly collected in our field. This may lead to the identification and measurement of new constructs. It may also help avoid the persistent faking and impression management so common in data collected for personnel selection (such as when measuring personality). For example, it has been noted that recruiters may review a candidate's social media (such as Facebook) because it represents a more honest presentation of the candidate's true personality compared to interview responses (Hartwell, & Campion, 2019).

However, the utility of data not collected specifically for CATA is likely to be limited by not directly focusing on the constructs of interest. To maximize utility of CATA-specific data collection, we should not rely just on the text data we normally collect (like asking for "any other comments"), but researchers should instead purposively collect text data to measure the intended or desired constructs. That is, if information is sought on a specific attribute, the questions should ask specifically about that attribute rather than inferring it from information collected for other purposes. For example, if the goal is to measure past leadership accomplishments, it is better to ask about past leadership accomplishments and text mine the responses than to text mine indirect text data such as personal statements. This is because the direct approach is more likely to solicit relevant and complete data on leadership, while the personal statement may focus on life goals and not discuss past accomplishments.

In summary, CATA can be used to analyze virtually any type of textual data. More than half the time past research has examined text data that was not collected for the study but was generated for other reasons. This fact, combined with the utility of CATA-specific data collections, suggests CATA may be able to identify new constructs because it is tapping into new data sources while also avoiding some of the problems with purposefully collected data (like impression management). However, if text data are collected to measure a particular attribute, the questions should ask about that attribute as opposed to inferring it from other indirect information.

3.3 Types of Analyses

The review of the types of text analyses used in the literature led to the development of a typology for organizing CATA approaches. Our typology was informed by the typologies used by Banks, Woznyj, Wesslen, and Ross (2018) and Medhat, Hassan, and Korashy (2014), but is necessarily different in most ways for two reasons. First, the existing typologies had a slightly different focus. Banks et al. (2018) was focused on the uses of R for text analysis, and Medhat et al. (2014) was focused on sentiment analysis and had more of an engineering context. We needed a model focused on measuring human constructs in a management context. Second, there is some obvious confusion in the literature on distinctions among types of content analyses that we wanted to avoid, as explained below.

A critical point of confusion is the terminology. Despite the long history of content analysis, there are several frameworks that attempt to organize the various approaches, and we have found that these do not always align. One recognized authority on content analyses is the book by Drisko and Maschi (2016). They define three types of content analysis: (a) Basic content analysis using word counts and other clustering analytic methods to derive categories using either deductive or inductive methods (Drisko & Maschi, 2016, p. 3), (b) interpretive content analysis using researcher-generated summaries and interpretations rather than word counts or other quantitative analytic methods inductively (p. 5), and (c) qualitative content analysis involving both inductive identification of categories and deductive application of these categories to additional data in an iterative process (p. 6). Other scholars writing on content analysis also make distinctions between methods. For example, Hsieh and Shannon (2005) refer to summative, directed, and conventional, the use of which depends on the state of the theory. These do not align directly with Drisko and Maschi (2016), but there are similarities. For example, Hsieh and Shannon's (2005) "summative" content analysis appears to be equivalent to Drisko and Maschi's (2016) "qualitative" content analysis rather than their "summaries." Similarly, Krippendorff (2004) distinguishes between approaches that are inductive with the goal of identifying and summarizing themes and those that are more deductive or driven by research questions to find information to solve a problem. Krippendorff also distinguishes between interpretive versus quantitative approaches.

The difficulty is that the definitions of inductive-deductive and summative-interpretive and the relationship between them in content analysis is not clear based on the literature. In an attempt to reconcile the different definitions, we would propose the following. Inductive and deductive are about how you develop your explanation or theory. Inductive derives the explanation or theory from the data, while deductive tests your explanation or theory based on whether predictions are supported by the data. In content analysis, summative means simply summarizing the data in some way. It is simply descriptive. While interpretive means using the data to come to some conclusion or answer some question that usually is known in advance and you are looking for confirmation. It is actively searching for something as opposed to passively describing. Summative studies are more likely to be inductive if they derive an explanation or involve theory development, but they might not involve induction if they are simply descriptive and not trying to develop a theory or explanation. Logically, however, summative studies are unlikely to use deduction because they are not trying to test theories or explanations. Interpretive studies are more likely to be deductive, but they might include induction (as in grounded theory studies that go back and forth).

Krippendorff (2004) provides a potentially relevant insight. Although inductive and deductive inferences are useful for content analysis, depending on the purpose, he argues that "deductive and inductive inferences are not central to content analysis" (page 36), but instead the inferences in content analysis tend to be better described as "abductive." Krippendorff says that "abductive inferences proceed across logically distinct domains, from particulars of one kind to particulars of another kind. (These are the kinds of inferences of interest to content analysis, where they proceed from texts to the answers to the analyst's questions.)" (page 36). He further explains that "abduction starts with a body of data (facts, observations, givens)—our text. A hypothesis—our analytical construct—if true, would explain these data. No other hypothesis can explain the data as well as the chosen one does. Therefore, the hypothesis is probably true and can be used to deduce other entailments—that is, answer our research questions (page 37)." In plain language, abductive inferences essentially mean inferring from an incomplete observation to a best prediction. Wikipedia defines abductive reasoning as "a form of logical inference which starts with an observation or set of observations then seeks to find the simplest and most likely explanation for the observations. This process, unlike deductive reasoning, yields a plausible conclusion but does not positively verify it" (https://en.wikipedia.org/wiki/Abductive_reasoning). For example, inferring a candidate's personality from past job types would be an abductive inference.

Given all this confusion, and the fact that every type of inference is useful in different contexts, we believe a more helpful distinction for guiding researchers on the choice of analysis is the amount of human intervention required. Some require little intervention, such as the use of an existing dictionary, while others require more intervention, such as developing categories. Moreover, none of these previous distinctions between content validity approaches recognize the more technically complex approaches, such as machine learning. Content validity approaches vary widely in degree of automation. Therefore, we also distinguish between the amounts of automation.

Our typology is shown in Table 2. It depicts four types of CATA, distinguished along two dimensions: (1) Whether the analysis required high or low human intervention and (2) whether the analysis relied on low or high computer automation.

The first dimension considers the amount of human intervention (low vs. high). Human intervention occurs in all types of CATA still, but to varying degrees and at different stages in the analytic process. For example, beginning in the top left cell, dictionary-based CATA can require relatively little human intervention. In this type of analysis, researchers can use either pre-built dictionaries or can develop their own. Even the task of developing their own is not cumbersome, particularly in comparison to categorization-based CATA. Significant human intervention is required throughout categorization-based CATA wherein researchers are actively reading excerpts, developing codes, and imposing meaning on the data through the codes. Interpretation of analysis occurs throughout categorization-based CATA, whereas outputs with dictionary-based analyses are fairly straightforward. For example, if we were interested in analyzing the sentiment of an open-ended response, the output would simply provide a count of the number of positive and negative words used in the response. Similar to dictionary-based CATA, unsupervised machine learning requires no human intervention during the analysis, though human judgment is still required to interpret the output. In contrast, supervised machine learning requires significant human supervision to ensure that correct synonyms are being identified and similar categories are being combined as the human trains the model.

The second dimension addresses the degree to which the analysis is automated (low vs. high). While it is quite obvious that unsupervised machine learning falls under high automation, perhaps the placement of dictionary-based CATA under low automation is less clear. It is easiest to think of dictionary-based CATA as essentially a word search. Much like searching a document in Word, dictionary-based software searches the text for words in the dictionary and assigns a "1" if they exist in the text and a "0" if they do not. Compared to machine learning, it is relatively rudimentary automation. Similarly, categorization-based CATA requires little automation. In these instances, automation only occurs in category management. For example, a researcher analyzing a series of interviews using Nvivo will save and color-code the categories in Nvivo and then simply click on other words or phrases to assign them to the categories. In contrast, machine learning is highly automated by the use of complex statistical procedures such as LSA and NLP.

Note that these types often build on one another in a study, sometimes moving toward more automation, such as categorization- or dictionary-based analyses informing supervised machine learning.

Table 2: Typology of CATA Approaches Based on Level of Human Intervention and Degree of Automation

	Low Automation	**High Automation**
Low Human Intervention	Dictionary-Based CATA	Unsupervised Machine Learning
High Human Intervention	Categorization-Based CATA	Supervised Machine Learning

Each of the four types of analyses (cells) in the typology can be used to identify constructs inductively (from the data with relatively little theoretical influence) or deductively (e.g., developed *a priori* or based entirely on some theory). As explained earlier, inductive analyses are conducted with little or no premonition as to the types of constructs that may emerge or the nature of the relationships among constructs. Researchers remain as close to the data as possible and allow constructs to emerge organically from that data and develop their own coding scheme. This is the traditional route to inductive coding. Advances in CATA use the computer as a tool to do the inductive coding, much like we did with Air Force Remotely Piloted Aircraft (RPA) Sensor Operator (SO) data. In contrast to inductive analyses, CATA also can be used for deductive analyses where constructs are identified *a priori* and/or by using theory and then sought out in the text and often quantified. Considering theory and previous research and then using that iteratively with coding the data involves both inductive and deductive components, as often occurs in grounded theory development studies as described previously. As will be described below, different CATA approaches illustrate the relatively more inductive to the relatively more deductive approaches, although probably all approaches use some of each.

Similarly, each of the four types of analyses (cells) in the typology can be used in either a summative or interpretive content analysis. Each approach could be used to summarize the data to simply describe it, and each could be used to interpret the data to come to a conclusion or answer a question that is known in advance.

Each of the four types of analyses (cells) in the typology can be used to analyze content or sentiment, and as such, it is important to distinguish between the two. Whereas content analysis uncovers *what is being said*, sentiment analysis uncovers *how it is being said*, usually determined by the affective directionality of the text (e.g., positive or negative). Broadly speaking, *content analysis* is a "research technique that uses a set of procedures to classify or categorize communications to permit valid inferences to be drawn" (Morris, 1994, p. 903). As a highly intuitive approach to understanding data, content analysis reduces comments (words) to digestible and coherent categories (variables) related to the ideas communicated in the text. This can be done at any level (word, phrase, line, sentence, document, etc.), allowing for multiple content categories to co-occur. *Sentiment analysis*, alternatively, is the assessment of "a view or attitude toward a situation or event; an opinion" (Dictionary.com, 2019, n.p.). Speer (2018) notes it also represents the "valence or intensity of a text" (p. 308). Sentiment analysis analyzes data on one of three levels of classification: Document, sentence, or aspect (Medhat et al., 2014). As sentiments occur in response to an event or situation, the number of sentiments in a text tends to depend on the number of events or situations to which the writer is responding. For example, a news article summarizing a series of organizational events (e.g., mergers, changes in stock prices, scandals) will likely include a sentiment for each event. Conversely, a news article reporting on a comment from a CEO will likely only include one sentiment. Looking across all the studies in our review, content analyses comprise two-thirds of the studies (146; 66.7%), whereas sentiment analyses comprise nearly a fourth of the studies (55; 25.1%), and several studies conducted both content and sentiment analyses (18; 8.2%).

3.3.1 Low Automation CATA

With *dictionary-based CATA*, the "occurrence of specific words from an existing list indicates the salience of a construct" (Short et al., 2018, p. 420). Almost half of the studies used this approach (109; 49.8%). They require a low amount of automation because they rely largely on basic word searches that yield simple word counts. They also require low human intervention because they commonly use an existing dictionary and, even if they develop one, it requires relatively less effort than the other approaches because of its simplicity. As noted above, of the four types of analyses, sentiment analyses are overwhelmingly conducted using dictionary-based approaches, though some content analyses utilize this method as well. For example, Gelfand et al. (2015) used LIWC (a dictionary-based tool generally used to detect sentiment) to evaluate linguistic features in negotiations. Dictionary-based CATA can be used for both sentiment and content, but most of the articles that used the dictionary-based approach in our review did so to assess sentiment. Dictionary-based CATA tends to be more deductive if it uses existing dictionaries and more inductive if it develops a dictionary, but some studies can be both.

The dictionary-based approach (using sentiment as an example) ordinarily goes as follows: (1) Scholars develop a dictionary of terms that represent the sentiment they are interested in mining, so if they are interested in both positive and negative sentiments, they would develop two dictionaries; (2) assuming the dictionary is validated, they then put the dictionary into the appropriate software (as an illustration of validation, see the report on LIWC in Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007); (3) researchers run sentiment analysis on their text data using the dictionary; and (4) the software counts the number of times a word (or its stem) from the dictionary is used in each response and assigns that count as the appropriate type of sentiment to that response. In most cases, dictionary-based analyses are used to quantify the text data, often to test hypotheses deduced from a theory (e.g., Brett et al., 2007; Gomulya, Wong, Ormiston, & Boeker, 2017). Because this count is affected by whether the participants are particularly wordy in their response, researchers can control for the total number of words participants use. It should be noted that some software (LIWC, DICTION) have pre-loaded dictionaries for particular content areas. Although dictionaries are frequently used with sentiment analysis, they can also be used for content analysis. In those cases, the researcher develops a dictionary based on words reflecting the categories of interest, and then uses the dictionary to content analyze documents in the same manner as above. The output would be the number of times documents used words reflecting the content categories, as opposed to the number of times the documents used words reflecting the sentiments. The level of automation is again simple because the computer only counts words.

Categorization-based CATA refers to approaches where codes and categorizations emerge from the text data with or without assistance from previous codes. Of the 219 non-review studies, 96 (43.8%) used a categorization-based approach. This approach usually only requires a low level of automation. For example, researchers often use simple approaches to content analysis, such as word counts (including visual techniques such as word clouds). When humans generate the list of codes or words, simple software is often used only to house and manage their data.

However, this type can also include the use of computers to generate the codes (e.g., extract concepts) and human judgment is then used to organize the codes, as well as to clean the codes such as eliminating “stop words” (e.g., the, a, an, in, etc.) or words that are so frequent they do not have meaning (e.g., “mission” in the Air Force Realistic Job Preview (RJP) data). Nevertheless, the automation is relatively less than some other approaches. Categorization-based CATA requires a relatively higher amount of human intervention because the categorizations usually rely on human judgment to a greater amount than some other approaches. This includes both creating the categories and interpreting their meaning. Finally, this approach can be either inductive or deductive. For example, some researchers may use a “seed list” of words as a place to begin (Medhat et al., 2014), even when the study is primarily inductive. The codes can be informed by prior theory or coding categories in more deductive studies. They can also be used as building blocks of theory (as in grounded theory studies) or can be quantified in some way relevant to the research question or hypothesis at hand (e.g., Belderbos, Grabowska, Leten, Kelchtermans, & Ugur, 2017). Purely qualitative studies (as defined in Table 1) are categorization-based CATA. All other types, and sometimes categorization-based CATA, have some quantitative component.

An advantage of the categorization-based approach over the dictionary-based approach is that it considers contextual text elements as relevant to opinion generation (Medhat et al., 2014). Applying a list developed a priori, as is the dictionary-based approach, may miss additional terms or phrases that represent key content or sentiment in that context. Organizational members often create their own language to increase efficiencies or as a symbol of their organizational culture and these go uncategorized in dictionary-based approaches where researchers use existing dictionaries, such as those already loaded in DICTION or LIWC. A good example of this is use of acronyms in the military. These acronyms harbor important meaning that will not be recognized with a dictionary-based CATA approach without the intervention of humans making modifications to the dictionaries. On the other hand, an advantage the dictionary-based approach offers is an already validated list, reducing the burden on the researchers to develop and validate their own list, as is the case with the categorization-based approach. Plus it is more objective and comparable across research studies.

3.3.2 High Automation CATA

The CATA approaches with higher levels of automation use various forms of *machine learning*. We provide two definitions of machine learning. The first is “Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use in order to perform a specific task effectively without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence” (“Machine Learning”, 2019, n.p.). The second definition provided by a developer attempts a similar definition with more brevity: “Machine Learning is the training of a model from data that generalizes a decision against a performance measure” (Brownlee, 2013, n.p.). Regardless of definition complexity, the main goal of machine learning is to eliminate human intervention and allow the computer to learn how to make or predict human-like decisions or discover new relationships among phenomena. This can save substantial time for the researcher and may lead to the discovery of new insights.

ML includes a wide range of techniques. The least sophisticated are techniques designed to simplify the data, which are often necessary because of the large number of variables. Common examples include clustering, which is an analysis that identifies similarities among data points in terms of distance and creates groups of observations based on those similarities, and dimensionality reduction, which occurs through feature extraction (such as reducing the number of dimensions using techniques such as principal components analysis) and feature selection (such as selecting only relevant variables and eliminating irrelevant variables).

Much more sophisticated techniques for machine learning use computer software to derive meaning from the text by applying NLP. NLP is the process through which a human "helps computers understand, interpret and manipulate human language" (SAS, 2019, n.p.). NLP is also sometimes called computational linguistics and researchers will discuss the use of "n-grams" (sequences of words in a string of text, with "n" being the number of words). It considers relationships among words as well as the words themselves. Critical to machine learning, NLP is the most commonly used highly sophisticated (automated) technique to use computers to understand and generate meaning from textual data. For example, in a study using 52,392 U.S. Securities and Exchange Commission (SEC) 10-K Annual Reports, Menon, Choi, and Tabakovic (2018) used a software called "Natural Language Toolkit" to build Python code to capture strategic change, strategic positioning, and strategic focus. NLP will also preprocess text by stemming (reducing words to basic forms: "playing" and "played" become "play") and lemmatizes (reducing words to dictionary form, such as converting "is" and "am" to "be") so the data can be more easily analyzed. Stemming and lemmatization are two of the first steps any NLP software takes because doing so allows for words to be recognized as the same across text. Whereas one piece of text might use the word "writing" and another uses the word "wrote," the stemming and lemmatization process reduces both to "write" indicating that each piece of text is referring to the action of writing. While it is possible to do this by hand, which was required years ago, scholars now have access to a number of new software programs that do this for them. Other similarly sophisticated machine learning techniques, such as LSA will be described later in this report.

As a final example of ML techniques, researchers might use data to train the computer. Although data are used to identify the coding categories, which is using data to train, training here refers to using relationships between the text categories extracted and various outcomes to select the categories to retain in the model. For example, Campion et al. (2016) extracted 5,000 text categories initially for each model in their data, but then used correlations with human raters to select only the 1,000 or so most predictive categories to retain in the model. To train the computer, researchers break their data into two sets – training and testing data. They use the training data to train the computer to identify the classifications, distinguish between groups, or predict outcomes, and then they use the remaining testing data to analyze whether the model accurately predicted. Not only are the data used to train the computer, but the researcher might use existing word dictionaries or might modify the variables extracted based on their degree of predictiveness, theory, or parsimony.

The criterion for success in studies that use human scores to train the computer is the resulting correlation between the human and computer scores. The size of the correlation will depend on how much effort is expended training the model, how successful the training is, how much shrinkage occurs in cross-validation, and possibly other factors (e.g., erosion over time). The maximum correlation is also limited by the reliability of the criterion measure (interrater reliability in this case) based on a well-known formula. The formula demonstrates that the maximum correlation with another variable is limited by the square root of the product of their reliabilities. So, assuming a perfect reliability for the computer score, a reliability of .60 for the human scores would limit the human-computer correlation to .77. Note also that the reliability of the human scores can be increased by training or by creating idealized examples of text at various score levels to train the computer.

Within these more automated approaches, the distinction between unsupervised and supervised approaches reflects the amount of human intervention dimension. Relatively *unsupervised machine learning* is exactly as it seems – the computer identifies and learns the structure of the data with little human intervention. The data are not labeled and the researcher provides little to no guidance to the computer as to how data might be reduced. It makes the assumption that no "true" categories or clusters exist and generally humans do not alter the groups the model identifies (hence, it is "unsupervised"). As such, it is completely inductive. Only two studies used this approach (two studies, .9%) in the management literature reviewed. It is likely to be much more common in other disciplines (information technology (IT) and engineering) and practical applications as opposed to journal publications.

Alternatively, *supervised machine learning* was somewhat more common in our review (12 studies, 5.5%). Supervised means it involves one or more of the many possible ways that the researcher guides the analysis. As such, we consider it to be relatively high on human intervention. Essentially, humans only participate in unsupervised learning when it is time to interpret the output, whereas in supervised learning humans participate at many different stages to guide the model (e.g., identifying synonyms, combining categories, etc.). Supervised ML can be used in both inductive and deductive studies, while unsupervised ML seems likely to be somewhat more common in inductive studies.

Of the example ML techniques described above, simplifying the number of variables using cluster or factors analysis would be considered fairly unsupervised because the researcher makes relatively little modification and the process is mostly inductive. However, when data are used to train a computer algorithm that is considered supervised ML not only for the reasons described above, but also because the researcher identified or collected the data with some purpose or theory in mind. Of course, anytime the researcher modifies the variables directly (such as telling the computer synonyms or picking the best variables) that would also be considered supervised.

One advantage of unsupervised ML, as Janasik, Honkela, and Bruun (2009) write, is that "unsupervised learning methods use input data to generate the output, the methods are not, in principle, vulnerable to researcher bias or *a priori* categorizations" (p. 443). On the other hand, some degree of supervision is usually required to improve the interpretability of the findings and relate them to the purpose (or theory) underlying the study.

We would like to note that through the development of machine learning in data science, programming, and other non-management or non-psychology contexts, the definition of and distinction between unsupervised versus supervised machine learning is, at times, imprecise and emerging. For the purposes of scientific research, we tend to think of it more along a continuum (from relatively unsupervised to relatively more supervised) than absolutes. We believe our distinction is reasonable and reflects the central difference of importance.

Appendix B includes a list of videos, websites, and podcasts on this topic that we have found helpful in the past and continue to use to keep up with the changing landscape of ML.

In summary, the types of CATA in the literature can be described in terms of a typology consisting of two basic dimensions – low versus high human intervention and low versus high computer automation – that define four types: dictionary-based, categorization-based, unsupervised ML, and supervised ML, with dictionary including sentiment and content analysis. Together, these describe the range of CATA used in the literature for content analysis as well as the range of choices available to researchers. Historically, the CATA in the literature has used mostly the simpler and less sophisticated approaches, but that is likely to change as researchers recognize the value of the more sophisticated approaches, the software becomes more accessible, and more information on how to use ML becomes available.

3.4 Types of Construct Validity Methods

There are several recognized authorities on the concept of validity when it comes to scores on tests or other measurements for the purpose of employee decision making. They include:

> *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice, 1978).
>
> Principles for the Validation and Use of Personnel Selection Procedures (Society for Industrial and Organizational Psychology, 2018).
>
> *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

The generally accepted scientific definition of validity is the "unitary" concept, which is that, although there can be different types of evidence for validity, it is one thing—the inferences that can be made from test scores (American Educational Research Association (AERA) et al., 2014, p. 11; Sheltered Instruction Observation Protocol (SIOP), 2018, p. 5). Moreover, because all types of validity bear on the meaning of the construct being measured, they are really all types of construct validity evidence (SIOP, 2018, p. 14). Therefore, we will refer to all of these types of evidence as informing construct validity, even though that term is often used to refer to correlations with other tests as we explain below. Our review shows that researchers have used all types of evidence to interpret the construct validity of CATA. We found it useful to distinguish between five types of validity evidence to best interpret this literature:

1. Evidence based on clearly defining the content domain. This is considered content-based evidence of validity by all the testing authorities (AERA et al., 2014, no. 1.11; Equal Employment Opportunity Commission (EEOC) et al., 1978, Section 14C4; SIOP, 2018, p. 28). It may include defining the domain theoretically in academic research studies, using job analysis in applied personnel selection, and explicitly generating item content to measure or sample the domain in all contexts. All of the studies in the literature review included this type of evidence in one form or another, thus we did not code it separately because it did not distinguish the studies. Moreover, the Air Force is familiar with this step because it is always a baseline in any measurement development study.

2. Evidence based on using SME judgments to interpret categories. This is a central component to content validity in the SIOP (2018, pp. 27-28). Studies in our review varied on whether they used this type of evidence to inform construct validity.

3. Evidence based on the internal structure of the measures. This type of evidence refers to statistical indicators of proper measurement and may include reliability or other psychometric statistical quality indicators, factor structure, or other information about the relationships among items or sections of the measure (AERA et al., 2014, no. 1.13; EEOC et al., 1978, Section 14C5; SIOP, 2014, pp. 32-33). Studies on CATA varied widely on whether this information was used to support construct validity and what types of statistical indicators were used.

4. Evidence based on relationships with other measures. Although historically this type of evidence has been called "construct" validity (EEOC et al., 1978, Section 14D), it is described in this broader way in the profession these days (AERA et al., 2014, no. 1.16; SIOP, 2018, p. 14). Studies of CATA varied widely in their use of relationships with other measures to interpret construct validity.

5. Evidence based on predicting outcomes or consequences. This type of evidence is often called "criterion-related" validity (EEOC et al., 1978, Section 14B), but we use this broader label to be consist with the professional description these days (AERA et al., 2014, no. 1.17; SIOP, 2018, p. 8). Again, studies of CATA varied widely in their use of criteria to interpret construct validity, so coding of this topic may be informative to the Air Force purposes.

Below we will discuss each type of method (except #1) and present examples of good practices of each from our review of the literature.

SMEs have been central to historical content validity efforts in psychology research. SMEs are individuals who have expertise in the area germane to the research topic at hand. For example, when conducting job analyses, researchers often gather input from job incumbents, supervisors, and human resource analysts to provide a well-rounded perspective on what the job entails. As such, it is not surprising this tried-and-true method remains a relevant procedure in the construct validation of categories derived with computer assistance. Nearly half (92; 42.2%) of the studies in our review employed SMEs. In some instances, one or more co-authors were left out of coding and used as SMEs (Kreiner, Hollensbe, & Sheep, 2006), while others used independent coders—such as other scholars or PhD students familiar with the research topic (Nadkarni & Chen, 2014). Finally, others used industry SMEs such as informants (or individuals who were interviewed; called "member checks" in grounded theory research) (Lingo & O'Mahoney, 2010; Wilhelmy, Kleinmann, Konig, Melcerhs, & Truxillo, 2016). In all, the use of SMEs to help support content validity remains as important to CATA as it has traditionally been to non-computerized content analysis.

The most obvious evidence based on the internal structure of the measures was to calculate psychometric quality estimates on the variables they created using CATA, which about a quarter of the studies did (58; 26.6%). Generally, these were estimates of agreement because the interest was in the reproducibility of the categories. Many types of agreement indices were used in the literature such as Cohen's kappa, Krippendorff's alpha, simple percent agreements, and Kendall's W. Traditional measures of reliability (defined as covariation as opposed to agreement), have been much less common. For example, intraclass correlation could be used to examine the reliability of coders (e.g., Colquitt, Long, Rodell, & Halvorsen-Ganepola, 2015). Our view is that agreement and covariation are different and important, and both should be calculated.

Another common type of evidence based on internal structure is the factor structure. Once the text-mined variables (concept categories) are extracted, they are often large in number, so reducing their dimensionality is desirable and can be achieved using a variety of methods like factor analysis and cluster analysis. Parallel to the discussion above, dimension reduction techniques can be based on covariation (like factor analysis) or agreement (defined as distance in n-dimensional space like cluster analysis). For example, Sbalchiero and Tuzzi (2016) utilized cluster analysis to assess the occurrence and co-occurrence of words in transcribed interviews to classify texts into groups called "lexical worlds" (Reinert, 1990) or classes based on similar word use. In another example, Hajek (2018) applied two dimensionality reduction techniques to simplify the feature space, including correlation-based feature selection where features with low correlations were removed because low correlations suggested low relevance, and LSA where the number of factors in the feature space were reduced using cosine similarity.

Although we did not see this in the literature, once composites are created based on dimensionality reduction, it would be reasonable to evaluate their internal consistency reliability. It would not only index the homogeneity of the composite, but it would have the other appealing interpretations of alpha, such as estimating the correlation between the scores obtained and the scores that would be obtained with another sample of text-mined variables drawn from the same domain. In other words, it would show that the specific set of variables extracted would not cause a large change in the scores, which is important because many researchers worry that the results of text-mining might not replicate.

Another common method researchers used to support the construct validity of their text analyses was to examine convergent and discriminant validity correlations with other measures, which 40 (18.3%) of the 218 empirical studies did. A number of techniques were used. Some offered convergent and discriminant validity using simple correlations with theoretically related and unrelated constructs (or regression; Chatman, Caldwell, O'Reilly, & Doerr, 2014). For example, Madera, Hebl, and Martin (2009) used independent coders to rate the descriptions of "agency" and "communal" and correlated these with the CATA measures of "agency" and "communal." Mossholder, Settoon, Harris, and Amenakis (1995) correlated the results from the Dictionary of Affect in Language (DAL) with organizational commitment, role ambiguity, and role conflict to support that the affect score using the dictionary sufficiently captured affect. Finally, Godnov and Redek (2018) correlated sentiment scores of written online ratings of hotels and spas with the satisfaction score also provided by the raters (often communicated via number of stars out of five).

The final method of supporting construct validity of CATA variables in the literature is using them to predict outcomes or using them as the outcome. Nearly half of the studies in our review (101; 46.3%) took this step. In most cases, CATA variables were used to predict other variables. For example, Vagnani (2015) assessed whether a firm's orientation toward exploration predicted its long-term performance. Orientation toward exploration was defined as "search, variation, risk taking, experimentation, play, flexibility, discovery, innovation" (March, 1991, p. 71). Using annual reports as the primary text to analyze, Vagnani applied CATA (did not disclose software) to build a dictionary to identify exploration-oriented words. He found that exploration predicted long-term organizational performance. In a more micro example using experiments, King, Shapiro, Hebl, Singletary, and Turner (2006) found negative emotionality of language was related to discrimination in customer service. As a final example, rather than using SMEs to conduct the content analysis as discussed above (e.g., deciding which text comments should be categorized together), some studies used SME ratings as the criterion against which the machine learning algorithm was trained (e.g., Abulaish, Jahiruddin, & Bharadwaj, 2019; Campion et al., 2016). These studies had SMEs rate the text on dimensions of relevance to the study (e.g., amount of skill described in the text), and then selected variables for the model based on correlations between the text mined variables (extracted using some automated technique) and the SME ratings. This is a useful methodology because relevant criteria are often not available, so using SMEs to create criteria is a viable alternative.

In other instances, researchers used the CATA variables as dependent variables, such as Lanaj, Foulk, and Erez (2019) who found that leader self-direction was indirectly related to leader "clout" through leader depletion and work engagement. Clout is the confidence with which a leader spoke as measured with LIWC using open-ended responses from leaders. Finally, some researchers use the CATA variables as controls. In their study on how CEOs use metaphors in communications, Konig, Mammen, Luger, Fehn, and Enders (2018) used LIWC to identify four of six types of CEO communication to control for them. The four types identified by LIWC were future orientation, image-based language, CEO's use of numerical language, and optimism of the CEO's tone. As such, CATA serves several important purposes in predictive analyses and should not be overlooked as a method to identify confounding variables.

Finally, it should be noted that studies can extract the text-mined categories and then evaluate their predictiveness of outcomes, as in the studies above, or studies can develop the computer model based on extracting and retaining the variables based on their prediction of outcomes. The former has been the most popular approach in the management literature because identifying the content categories was the primary goal of the study (e.g., based on theory) and the correlations with outcomes was important validating information. However, the latter may be more useful in other instances where the goal is to predict an important outcome. This has been much more common in research studies in other areas outside organizational behavior (e.g., predicting consumer choices; Abulaish, Jahiruddin, & Bhardwaj, 2019; Bilro, Loureiro, & Guerreiro, 2019). However, accuracy of prediction is a preeminent goal when developing systems to make decisions regarding human resources, so the latter approach may be preferred (e.g., Campion et al., 2016). Of course, the two approaches are not mutually exclusive. For example, perhaps variables can be extracted based on their predictive value, but then other approaches to construct validity can be used to interpret the meaning of the variables.

3.5 Types of Software

3.5.1 Description of Software in the Literature Review

Types of software and the frequency with which they were used in studies in the review are listed in Table 1. The list goes from most common to least common. Only software used twice or more were listed for parsimony. The remaining software used only once in the review are grouped as "Only Used Once (Not Listed Individually Here)."

The most commonly used types of software for CATA are those developed for the primary purpose of content analysis, or reducing large amounts of qualitative data into a digestible number of representative categories. Nvivo (60; 24.7%) and Atlas.ti (18; 7.4%) were used in 32.1% of the studies. We later discovered that Nvivo is the successor to Nud.ist, so they are basically the same. These programs allow researchers to engage in inductive coding, discussed in detail in the Type of Studies section above.

An equally popular software in our review was LIWC (56; 23.1%). Researchers used this program most commonly to assign scores to text based on the number of sentimental words (positive or negative) present in the text using validated dictionaries of words categorized by sentimental meaning. However, it can and has been used for content analysis purposes (e.g., Martin, 2016). This software was released to the public in 1993 and has since undergone several iterations to consider changes in language or cultural modifications of word sentiment. The development of the software occurred in several stages to validate it. Beginning with general word generation, researchers amassed a large database of words and resources (e.g., English dictionaries, Roget's Thesaurus, affects scales). Next, they solicited feedback from SMEs (or "Judges" as they are referred to in the manual) to provide their expertise regarding what should and should not be included in the dictionaries. Finally, they evaluated the psychometrics of the software, considering word use frequency and other elements. Information on its psychometric properties are readily available (Pennebaker et al., 2007). It can also be used to score personality and motivations. For example, Hirsh and Peterson (2009) found several linguistic correlates of the Big Five such that achievement-related words were positively associated with conscientiousness; anxiety and negative emotions were related to neuroticism; extraversion and agreeableness were both associated with family-related words or those that represented interpersonal concern; and openness was associated with words related to perceptual processing such as hearing and seeing. Though not used as frequently, SentiWordNet (5: 2.1%), AFINN Sentiment Lexicons (2; 0.8%), and Valence Aware Dictionary and sEntiment Reasoner (VADER) (2; 0.8%) have similar functions to LIWC.

DICTION (13; 5.4%) is another software program that uses built-in dictionaries to score text data. Much like LIWC, DICTION is a particularly flexible program that allows for both content and sentiment analyses. However, the built-in dictionaries were developed specifically to extract the following qualities: certainty, activity, optimism, realism, and commonality (DICTION, 2019). Yet, researchers are not limited to these attributes and can generate and use their own dictionary(ies) in DICTION. For example, in their study on leadership rhetoric, Bligh, Kohles, and Meindl (2004) used DICTION's optimism dictionary and developed five dictionaries in addition to optimism (collectives, faith, patriotism, aggression, and ambivalence) to determine attributes of leaders from presidential speeches. Finally, a less popular and less capable software program that uses dictionaries, or word lists, is MonoConc Pro (4; 1.7%). This software identifies and counts words from a pre-defined word list in textual data to measure whatever attribute the researcher is interested in and has a dictionary for.

The more sophisticated software used in our literature review include R (13; 5.4%), Python (including NLTK) (7; 2.9%), and SPSS Modeler (2; 0.8%), which allow researchers to apply advanced analytics such as LSA and NLP. However, 11 of the 13 R studies used R for only sentiment or content analysis and did not use these advanced analytics. LSA and NLP involve extracting the meaningfulness of narrative information by analyzing the relationships among multiple words. LSA assumes that words that are close in meaning will occur in similar pieces of text, and will use matrices of documents crossed with words to depict the mathematical relationships. This is used to create vectors representing the words in each passage. Relationships between vectors identify similar meanings. NLP uses machine learning to create rules to identify patterns among words to extract meaning, and to identify patterns between words and various outcomes, using probabilistic (statistical) models for both. The computer learns the rule by using statistical inference to find relationships based on a large sample of text data (or corpus). Latent Dirichlet Allocation (LDA) is used as part of NLP and is the same as probabilistic LSA where the distributions of relevant words in a document are assumed to be sparse. It can also be thought of in terms of document-word matrices and vectors.

A small, but notable proportion of studies (21; 8.7%) used software that no other researchers used in our review. With the exception of SAS (used once; Kakol, Nielek, & Wierzbicki, 2017), none of these software programs were recognized by the authors of this report as some were homemade by researchers or have become outdated with the advent of more advanced systems.

An important observation is that fewer studies than expected used the more sophisticated approaches incorporating LSA or NLP (52; 23.9%). Furthermore, few used commercial software products like SPSS (two studies) and SAS (one study). This is surprising because commercial products are more complete, have better documentation, and are more user friendly. This may be because such products are fairly new on the market or that they are still very expensive. Regardless, the important takeaway point here is that the software used by management scholars to this point has been the less sophisticated software. It relies on rather simplistic applications of CATA based largely on word counts and sentiment, with significant human judgment often required (especially for the content analysis). This may be, in turn, limiting the sophistication of the research possible using CATA. Using more advanced approaches may allow the researcher to derive more construct validity information and broaden the range of applications, which is the goal of the Air Force.

3.5.2 Information and Evaluation on the Most Common Software

It appears as though the Air Force has eight choices when it comes to software to perform CATA, which can be divided into four types: (a) content analysis software (Nvivo and Atlas), (b) word dictionaries/sentiment analyses software (LIWC and DICTION), (c) programming software / languages (R and Python), and (d) commercial predictive analytics packages that include text mining modules (SPSS and SAS). Therefore, we researched and evaluated those eight software packages. Table 3 below summarizes our findings in terms of each of the topics relevant to the Air Force.

Table 3: Content Analysis Software

Software Name	**Nvivo** (earlier version called “Nud.ist”, or Non-numerical unstructured data indexing, searching, and theorizing)	**Atlas.ti**
Source to acquire & cost	Access website https://www.qsrinternational.com/nvivo/home. Cost information https://www.qsrinternational.com/nvivo/products. Academic license is $700 (Pro) & $800 (Plus) Government license is $979 (Pro) & $1119 (Plus) Commercial $1399 (Pro) & $1599 (Plus)	Access website https://atlasti.com/. Cost information https://atlasti.com/product/licenses-prices/. Academic license for single user is $750 and increases incrementally depending on number of user licenses (e.g., $3,000 for 5; $5,7630 for 10); Government license for single user is $1,290 and increases incrementally depending on number of user licenses (e.g., $4,650 for 5; $9,300 for 10); Commercial license for single user is $1,840 and increases incrementally depending on number of user licenses (e.g., $6,800 for 5; $13,200 for 10)
Function	Generally used for inductive content analyses; humans generate their own codes for the text data and assign those codes to words, phrases, or lines of data; software keeps track of codes; not capable of more advanced modeling.	Generally used for inductive content analyses; humans generate their own codes for the text data and assign those codes to words, phrases, or lines of data; software keeps track of codes; not capable of more advanced modeling.
Information on efficacy (e.g., validity)	While it has the capability of extracting its own categories, this function is largely based on frequency of words and is less useful; otherwise, categories are entirely human-developed, and psychometric and SME agreement analyses are conducted outside the software.	Categories are entirely human-developed, and psychometric and SME agreement analyses are conducted outside the software.
Information on usability (e.g., difficulty to learn; technical support)	Relatively intuitive to use (used in ~25% of studies in our review); technical support appears readily available (https://www.qsrinternational.com/nvivo/support-overview) and there are online trainings on how to use the software (https://www.qsrinternational.com/nvivo/nvivo-training), as well as videos on YouTube.	Relatively intuitive to use and though only used in ~8% of studies in our review, it is a recognized software for inductive coding in management research; technical support appears readily available (https://atlasti.com/support/) and there are free and for-cost training available (https://atlasti.com/learning/), as well as videos on YouTube.
Likely applicability to RFP data	Can be used to analyze textual data of all types.	Can be used to analyze textual data of all types.
Likely applicability for measuring attributes in RFP	Because categories are developed by humans, this software is flexible for researchers to code whatever they want to measure.	Because categories are developed by humans, this software is flexible for researchers to code whatever they want to measure.
Anticipated advantages	Easy to learn, low cost, widely used, and flexible to meet researcher coding needs.	Easy to learn, low cost, widely used, and flexible to meet researcher coding needs.
Anticipated disadvantages or problems	Validity and reliability studies occur outside of software; not capable of more advanced modeling.	Validity and reliability studies occur outside of software; not capable of more advanced modeling.
Overall evaluation for potential usefulness	Does not appear to have capabilities beyond the commercial packages (described below), which also performs content analysis better. As such, if a commercial software was used, there would be no need for this software.	Does not appear to have capabilities beyond the commercial packages (described below), which also performs content analysis better. As such, if a commercial software was used, there would be no need for this software.
Word Dictionaries / Sentiment Analyses Software		

Software Name	**LIWC** (Linguistic Inquiry Word Count)	**DICTION**
Source to acquire & cost	Access website http://liwc.wpengine.com/. Cost information https://www.receptiviti.com/liwc-api-get-started. Academic license is $89.95; commercial version available, but must apply through Receptiviti. The manual also states that any commercial use of the LIWC dictionaries is forbidden without permission through Receptiviti (p. 20, https://liwc.wpengine.com/wp-content/uploads/2015/11/LIWC2015_OperatorManual.pdf)	Access website https://www.dictionsoftware.com. Cost information http://www.dictionsoftware.com/order/. Academic license is $219 and corporate license is $269.
Function	Originally used to analyze sentiment (affect) from text using in-house, pre-validated dictionaries. It has about 50 dictionaries in 7 topic categories of psychological processes as well as a number of other language metrics (e.g., number of types of words used). They are described in the manual. It can also be used to analyze content of text. Allows for the use of a researcher's own custom dictionary, but not capable of more advanced modeling.	Can be used for both content and sentiment; has five in-house dictionaries (certainty, activity, optimism, realism, and commonality; see more https://www.dictionsoftware.com/diction-overview/); Additional information on DICTION in attached "About DICTION" document; Allows for the use of a researcher's own custom dictionary, but not capable of more advanced modeling.
Information on efficacy (e.g., validity)	Fairly transparent in the development and refinement of their dictionary; Internal consistencies are in the attached document titled "LIWC2015_LanguageManual"	Not at all transparent about development and refinement of dictionaries.
Information on usability (e.g., difficulty to learn; technical support)	Due to its prevalence in our review (used in ~26% of studies) and its long-standing history (first released in 1993), it appears relatively easy to learn and use. Technical manual is attached ("LIWC2015_OperatorManual") and technical support is available via email by two of the researchers who developed the software (information in Operator Manual) as well as videos on YouTube.	Not as well-known as LIWC, but still recognized (~6% of studies in our review used DICTION). Technical manuals are attached ("Manual_DICTION" and "Using DICTION") and technical support is available via email or phone (information in technical manual) as well as videos on YouTube.
Likely applicability to RFP data	Can be used to analyze textual data of all types.	Can be used to analyze textual data of all types.
Likely applicability for measuring attributes in RFP data (e.g., leadership, whole airman concept, task knowledge/proficiency)	Because researchers are able to develop their own dictionaries, it is possible to use LIWC to measure attributes relevant to the purposes of this project. Should sentiment of narratives also be of interest, this software is capable of doing that without necessitating a new dictionary.	Because researchers are able to develop their own dictionaries, it is possible to use DICTION to measure attributes relevant to the purposes of this project. Should the five in-house dictionaries be of interest, this software is capable of doing that without necessitating a new dictionary.
Anticipated advantages	Straightforward, validated, and recognized software. Has a large number of in-house dictionaries.	Seemingly straightforward and recognized software.
Anticipated disadvantages or problems	Best at measuring sentiment only and not capable of doing more advanced machine learning procedures.	Not transparent about validity and reliability of dictionaries. Only five in-house dictionaries. Best at measuring sentiment only and not capable of doing more advanced machine learning procedures.
Overall evaluation for	May be useful if we want to measure sentiments. Because it cannot do more advanced machine learning,	May be useful if we want to measure sentiments that it contains. Because it cannot do more advanced machine

potential usefulness	its role will only be adding the sentiment analysis. Commercial packages may also have sentiment capabilities, but this is the best known.	learning, its role will only be adding the sentiment analysis. Commercial packages may also have sentiment capabilities.
Programming Software/Language		
Software Name	**R** (The R Project)	**Python**
Source to acquire & cost	Access website https://www.r-project.org/. It is free	Access website https://www.python.org. It is free Tutorial https://docs.python.org/3/tutorial/index.html.
Function	R is a highly powerful and flexible software. It is popular among data scientists and more recently (past 10 years or so) has been used by management scholars. It is open-sourced, meaning code is made freely available by other researchers. It can conduct traditional statistical analyses, content and sentiment analyses, text mine, and is capable of advanced modeling.	It is a programming language. It is highly powerful and flexible and can be used for myriad tasks including data analyses, text analysis and machine learning, and developing websites/mobile apps. Need to download packages to be able to run analyses. It is open-sourced and packages are available on Python's website, as well as other places online.
Information on efficacy (e.g., validity)	Can use pre-validated dictionaries to conduct analyses and validity and reliability analyses can be done using R.	Can use pre-validated dictionaries to conduct analyses and validity and reliability analyses can be done using Python.
Information on usability (e.g., difficulty to learn; technical support)	As it requires basic understanding of software coding, it may be difficult to learn. Because it is a free, open-sourced software, no person or team is available to answer questions. The first line on the software's help page is "Before asking others for help, it's generally a good idea for you to try to help yourself," and they provide a few ways in which researchers can do that. Otherwise, materials abound from books (recommended: Discovering Statistics Using R by Field et al. accessible https://www.discoveringstatistics.com/%20books/discovering-statistics-using-r/; Silge & Robinson accessible https://www.tidytextmining.com) to websites (e.g., Stack Overflow; GitHub, coding from NLP seminar from SIOP 2019 available https://github.com/coryamanda/SIOP2019_NLP_organization_Research), as well as videos on YouTube.	As it is a programming language and requires basic understanding of software coding. It may be difficult to learn. There are many guides online, as well as books, and there is an email for concerns not addressed on the site (https://www.python.org/about/help/). Many packages can be found on GitHub and additional directions can be found on YouTube.
Likely applicability to RFP data	Can be used to analyze textual data of all types.	Can be used to analyze textual data of all types.
Likely applicability for measuring attributes in RFP	Because researchers are able to develop their own measures, it is possible to use R to measure attributes relevant to the purposes of this project.	Because researchers are able to develop their own measures, it is possible to use Python to measure attributes relevant to the purposes of this project.
Anticipated advantages	Flexible and powerful and some coding available from others.	Flexible and powerful and some coding available from others.
Anticipated disadvantages or problems	Unless you understand the minutiae of software coding, you run the risk of conducting the wrong analyses, particularly if you simply use or repurpose someone else's coding. It will likely be time consuming to learn, especially for nonprogrammers.	Unless you understand the minutiae of software coding, you run the risk of conducting the wrong analyses, particularly if you simply use or repurpose someone else's coding. It will likely be time consuming to learn, especially for nonprogrammers.
Overall evaluation for	Could be very useful, but complexity and time to learn are likely to be prohibitive for the Air Force, especially	Could be very useful, but complexity and time to learn are likely to be prohibitive for the Air Force, especially

potential usefulness	when other options that can serve the same needs and will be easier to use are available.	when other options that can serve the same needs and will be easier to use are available.
Commercial Predictive Analytics Software Packages		
Software Name	**SPSS Modeler Premium**	**SAS Enterprise Miner**
Source to acquire & cost	Access website https://www.ibm.com/products/spss-modeler. Access costs https://www.ibm.com/products/spss-modeler/pricing. (also see differences among tiers) Free 30-day trial, $199/month subscription, $7,430 for Professional license, $12,400 for Premium license, and $25,600 for Gold, but Premium or Gold needed for text mining capability. Gold includes “Collaboration & Deployment Services for model deployment and management.”	Access Text Miner website https://www.sas.com/en_us/software/text-miner.html. Access Enterprise Miner website https://www.sas.com/en_us/software/enterprise-miner.html. Text Miner is a component of Enterprise Miner, which allows for machine learning. Seems to require two licenses with Enterprise Miner downloaded first. We are waiting to hear on pricing despite multiple requests.
Function	It is a user-friendly, powerful, and flexible software. It allows for more advanced features such as concept extraction and text mining, categorizations, and machine learning. It allows extensions to embed R and Python code into an SPSS modeler stream by simply inserting a “node” (analogous to a plugin). This allows running R and python scripts to import data, apply transformations, build and score models, display outputs, and export data. Plus, as part of a predictive analytics suite, it includes a very wide range of statistical procedures that are preprogrammed and easy to apply.	It is supposed to be a user-friendly, powerful, and flexible software. It allows for more advanced features such as concept extraction and text mining, categorizations, and machine learning. Plus, as part of a predictive analytics suite, it includes a very wide range of statistical procedures that are preprogrammed and easy to apply.
Information on efficacy (e.g., validity)	Has several built-in dictionaries, but there is little if any detail on the psychometric properties of the dictionaries because it is deemed “proprietary.” (see https://www.ibm.com/developerworks/community/forums/html/topic?id=1a202834-fb7c-4025-9bd3-f04707b992c4). Researchers are able to develop and validate their own dictionaries or add concepts to existing ones. Manual also discusses analyzing candidate sentiment using sentiment text link analysis, but unclear if include in-house sentiment dictionaries.	It appears SAS does not use a dictionary-based approach, but instead “relies primarily upon pattern recognition” (p. 10; see http://www.arnoldit.com/articles/sas-white-paper.pdf). It also appears that SAS Sentiment Analysis is a separate software (see http://support.sas.com/publishing/pubcat/chaps/65646.pdf).
Information on usability (e.g., difficulty to learn; technical support)	Similar to basic SPSS, it is a point-and-click software and therefore user friendly. As most researchers in I-O and management use SPSS, the interface will not require much acclimation. There are also many avenues of support including IBM technical support (see https://www.ibm.com/products/spss-modeler/support), videos on YouTube, and many other online resources.	Researchers in I-O and management who use SAS will not likely face much acclimation in regards to interface. There are many resources through SAS (see https://support.sas.com/documentation/onlinedoc/txtminer/whatsnew42.html), as well as videos on YouTube and other online guides.
Likely applicability to RFP data	Can be used to analyze textual data of all types. Also able to incorporate quantitative variables simultaneously using a wide range of built in predictive models.	Can be used to analyze textual data of all types. Also able to incorporate quantitative variables simultaneously using a wide range of built in predictive models.
Likely applicability for	Because researchers are able to develop their own measures, it is possible to use SPSS Modeler Premium	Because researchers are able to develop their own measures, it is possible to use SAS Text Miner to

measuring attributes in RFP	to measure attributes relevant to the purposes of this project.	measure attributes relevant to the purposes of this project
Anticipated advantages	Flexible and powerful, straightforward, familiarity, and support is available.	Flexible and powerful, straightforward, familiarity, and support is available.
Anticipated disadvantages or problems	Cost is greater than other software.	Cost is greater than other software.
Overall evaluation for potential usefulness	This is the software we adopted based on extensive consideration of all the alternatives. Due to this, we are the most familiar with it. This would be our clear recommendation if the Air Force can find a way to get it approved for download on its computers.	This was our second choice when we reviewed the alternative when we began working in the area because it was much more expensive. Since we did not adopt it, we are not familiar with it intimately.

NOTE: The request for proposals (RFP) discussed measurement of a broad range of attributes including skills, abilities, and other characteristics (e.g., writing skills, level of motivation for a given job or job type, personality, temperament, behavioral intentions.

4.0 CONCLUSIONS & RECOMMENDATIONS: ANSWERS TO QUESTIONS OF IMPORTANCE TO THE AIR FORCE

4.1 What Attributes Have Been Measured?

Researchers in different areas have focused on widely different topics, ranging from organizational behavior topics (e.g., attitudes and traits related to leadership, teamwork, voice, etc.) to topics in other areas of business (e.g., marketing, customer reactions, insurance claims, etc.). Table 4 lists the constructs that researchers intended to measure in the studies contained in the literature review. To derive Table 4, we first identified all the constructs in the literature, and then sorted them into categories of similar constructs. We found seven categories reflecting human attributes, one reflecting process relationships among the attributes, and one category containing unique attributes. They are explained briefly below, along with their potential relevance to the needs of the Air Force.

1. Sentiment: This is one of the two largest categories (47; 23.7%), undoubtedly due in part to the early and influential development of sentiment-scoring software (like LIWC), as described above. Although we originally questioned the value of sentiment scoring for personnel decisions, preferring to focus only on the content of candidate or employee information, we have identified several potential uses that may be relevant to the Air Force. First, employers want candidates with a "positive attitude." As every human resource (HR) professional knows, employees with a "negative attitude" are difficult to manage, hurt the morale of other employees, and create unnecessary workload for HR. Importantly, job satisfaction is consistent across jobs (Dormann & Zapf, 2001) and it is related to negative affectivity (Connolly & Viswesvaran, 2000). The difficulty with measuring attitude when hiring is that direct questions such as in an interview are vulnerable to faking. Using sentiment analysis to assess attitudes from written materials may be able to measure positive attitude in a way that makes response bias more difficult because it is less direct (less obvious to candidates) and it will be more difficult for them to know which word choices are scored. Second, the measurement of sentiment may be useful when analyzing survey responses (e.g., such as the surveys of new hires used to develop realistic job previews) because the favorableness of the comments might be more objectively measured with sentiment analysis than with human reader judgment. Third, for the same reason, sentiment analysis might be useful to analyze employee reactions to training because they could measure the tenor of the comments more objectively.

2. Cognition: This is the other of the two largest categories (about 50 studies or 25.3%). Although almost none of this research has been conducted in the context of personnel decision-making systems, cognitions are a category of attributes that are likely to be applicable. In addition to cognitions relevant to knowledge, skills, and abilities that might be relevant to hiring, cognitions around employee perceptions of organizations and their fit may be relevant to predicting and influencing turnover or reenlistment. To illustrate, a recent study in the *Journal of Applied Psychology* used machine learning to predict turnover among school teachers; many of the reasons for leaving included in the model were cognitions about the job (Sajjadiani, Sojourner, Kammeyer-Mueller, & Mykerezi, 2019).

3. Organizational Characteristics that Influence Stakeholder Psychological Reactions: This was the next largest category (about 26 studies or 13.1%). These attributes might also be relevant to predicting and influencing turnover or reenlistment because employees are often attracted or

repelled from an organization due to their psychological reaction to it (Ng, Yam, & Aguinis, 2019).

4. Behavior/Skill: This category was reflected in a fair number of studies (about 20 studies or 10.1%) with some of them directly relevant to staffing. More importantly, the attributes in this category could be expanded to include the attributes measured in staffing, given that many staffing tools rely on information about past behavior or the possession of skills.

5. Personality/Orientations: This category is also very common (about 16 studies or 8.1%), and these attributes are also directly applicable to hiring. Faking is a critical issue with personality tests. CATA-based measures of personality from open-ended questions about accomplishments should be much less fakable. This is because, unlike personality inventories, candidates do not directly score themselves, but describe themselves and the CATA scores their personality, similar to interview-based measures where the interviewer makes the judgements.

6. Language that Influences Human Psychological Reactions: There were a fair number of articles that did not measure human attributes directly, but instead reactions to the language used by organizations (about 17 studies or 8.6%). The usefulness of these studies for the current Air Force project are unclear, other than just the general awareness that CATA can be used to detect meaning in text beyond the content or the sentiment.

7. Leadership: Only five studies measured leadership attributes (2.0%), but such attributes have obvious relevance to hiring future leaders such as officer candidates. Although not examined in any of these studies, it might be possible to extract leadership attributes from textual material submitted by applicants.

8. Processes that Reflect Relationships among Attributes: Another five studies measured processes, defined broadly (2.0%). Measuring relationships among attributes might potentially have value in understanding phenomena that occur over time, such as the turnover/reenlistment process, but no research has examined that yet in the literature we reviewed.

9. Unique Attributes: These are 14 studies (7.1%) on separate topics that are categorized together here because they could not be categorized elsewhere. Most are unique “one-off” topics unrelated to the Air Force project.

Table 4: List of Intended Attributes Measured in the Literature

Sentiment

1. Affective component of celebrity of a firm; nonconforming language
2. Aggression and positive emotions
3. Announcement characteristics (positive or negative)
4. Causal words; positive emotions
5. Cognitions and emotions
6. Content and sentiment (2)

23. Neutral, happy, sad, and contemptuous*
24. Online customer engagement (sentiment)
25. Opinions
26. Opinions/behavior of firm in a proactive or reactive way referring to three institutional pillars (regulative, normative, and cognitive)

7. Customer knowledge on laptop brands, sentiment
8. Discrete emotions (hope, anger, and betrayal)
9. Emotion (2)
10. Emotion and cognition
11. Emotion helping
12. Favorability (CEO media tenor)
13. Favorable/unfavorable; trait/behavior focus
14. Five master variables from DICTION: certainty, optimism, activity, realism, and commonality
15. Future orientation; image-based language; use of numerical language; optimism of tone
16. Investor sentiment
17. Media favorability
18. Media reactions (positive and negative)
19. Narrative positivity and negativity
20. Narrative sentiment scores
21. Negative emotions*
22. Negative tone (press release tenor); Overall affective tone

27. Personality and sentiment*
28. Positive/negative affect of delegating tasks
29. Positive-negative sentiment*
30. Promotion and prevention words*
31. Reactive-affective conflict
32. Sentiment (20)*
33. Sentiment (anger)*
34. Sentiment (favorability) of media
35. Sentiment (positive, negative)*
36. Sentiment and attention
37. Sentiment and topic modeling
38. Sentiment differences in online customer reviews between English and Japanese reviews
39. Sentiment of media reaction to mergers and acquisitions
40. Sentiment of news article
41. Sentiment of reviews
42. Sentiment, content, and topics
43. Tenor of media coverage
44. Tone
45. Tone of discourse
46. Topic-based sentiment
47. Warm-glow rhetoric

Cognition

1. Attention patterns of TMT
2. Attention timing and attention intensity
3. Attentional homogeneity
4. Authenticity of multiple work identities

23. Identity
24. Identity and voluntary turnover
25. Identity resurrection
26. Identity work
27. Investor gut feel
28. Justification

5. Blame
6. Celebrity
7. CEO commitment to status quo (CSQ)
8. CEO regulatory focus*
9. CEO temporal focus
10. Changes in professional role identity
11. Clout
12. Cognitions and emotions
13. Cognitive meaning and values
14. Cognitive structure of cognitive ability (centrality and complexity)
15. Content credibility
16. Customer knowledge on laptop brands, sentiment
17. Decoupling
18. Depth of cognitive processing*
19. Dialectic tensions and organizational identity renegotiation
20. Emotion and cognition
21. Five master variables from DICTION: certainty, optimism, activity, realism, and commonality; Three calculated variables from DICTION: insistence, variety, embellishment, and complexity; Pearce & David's (1987) 8-item typology of what is included in a mission statement
22. Identify reconstruction
29. Justification content
30. Knowledge hiding
31. Level of ambivalence*
32. Misfit at work*
33. Moral foundations
34. Organizational identification
35. Organizational identity*
36. Organizational identity, language, and affiliations
37. Person-environment fit*
38. Psychological capital (2)*
39. Relational boundary work
40. Retrospective sensemaking of change to the work environment over time
41. Rivalry
42. Schemas
43. Sensemaking (2)*
44. Sensemaking of cultural distance
45. Spirituality
46. Split identification
47. Team cognitive maps*
48. Unearned status gain
49. Value maintenance work
50. Values*

Organizational Characteristics that Influence Stakeholder Psychological Reactions

1. Accounts of downsizing
2. Corporate Entrepreneurship
3. Corporate Social Responsibility (CSR)
4. Culture-specific problems and job attractiveness
17. Organizational exploration
18. Organizational separation of two or more autonomous organizational entities
19. Organizational values (2)

5. Enterprise strategy
6. Environment, risk management, reassurance, comparison, attribution, strategy, and people
7. High-stakes institutional translation
8. Institutional agency
9. Institutional voids
10. Lean professional competency taxonomy*
11. Legacy identification
12. Legitimacy
13. Logic hybridization
14. Management innovation*
15. Managerial discretion; sustainability strategies
16. Organizational culture: competitiveness, control- and coordination-orientation, customer-oriented, human-resources-oriented, innovation- and learning-orientation, and team-orientation
20. Personality: moderation, friendliness, intellectual brilliance, machiavellianism, poise and polish, achievement drive, forcefulness, wit, physical attractiveness, pettiness, tidiness, conservatism, inflexibility, and pacifism*
21. Recruitment signals and signal strength
22. Results-focused culture
23. Risk*
24. Stages of corporate translation of climate change: framing, localizing, normalizing
25. Stakeholder elements of the task environment
26. Strategic change, strategic positioning, and strategic focus

Behavior/Skill

1. Achievement*
2. Advocacy work; category work
3. Balancing micropractices
4. Boundary control efforts
5. Boundary management tactics
6. Communication skills, critical thinking, people skills, leadership skills, managerial skills, and factual knowledge*
7. Competition and cooperation
8. Conflict
9. Coordination process (2)
11. Interviewer impression management
12. Job search behaviors
13. Just and unjust treatment from supervisors
14. Knowledge practices*
15. Managerial discretion; sustainability strategies
16. Nexus work practices
17. Peer reference
18. Ritual performance
19. Tournament rituals

10. Creative use of resources

20. Work-family conflict and health

Personality/Orientations

1. Adaptability*
2. Agency
3. Big Five: Openness, conscientiousness, extraversion, agreeableness, neuroticism*
4. CEO narcissism*
5. Clock-time orientation
6. Communality and agency: communal adjectives, social-communal orientation, agentic adjectives, agentic orientation
7. Entrepreneur orientation (EO)*
8. Entrepreneurial orientation (autonomy, innovativeness, proactiveness, competitive aggressiveness, risk taking)*
9. Exploration orientation Extraversion*
10. Favorable/unfavorable; trait/behavior focus
11. Future orientation; image-based language; use of numerical language; optimism of tone
12. Global mindset
13. Market orientation; psychological capital
14. Need for achievement*
15. Personality and sentiment*
16. Top management team modesty

Language that Influence Human Psychological Reactions

1. CEO-CFO language style matching
2. Creativity-relevant language
3. Dialectic tensions and organizational identity renegotiation
4. Future orientation; image-based language; use of numerical language; optimism of tone
5. Inclusive and exclusive language
6. Linguistic signatures
7. Linguistic style*

11. Organizational identity, language, and affiliations
12. Rational and normative language
13. Rhetorical tactics to discern lifecycle of management research topic
14. Temporal cues (temporal framing: vagueness, distance)
15. Textual associative patterns
16. Rhetoric (inclusiveness, exclusiveness, tentativeness, certainty)
17. Vagueness (of wording)

8. Metaphors 9. No attributes, just language used 10. Objective language use (self-verification striving) and function words	
Leadership	
1. Humble leadership 2. Leadership development challenges 3. Leadership rhetoric: optimism, collectives, faith, patriotism, aggression, and ambivalence*	4. Leadership styles: symbolic, behavioral, political, and structural (and none of these)* 5. Leadership*
Processes that Reflect Relationship Among Attributes	
1. Dynamics of trace data 2. Dynamism of global teams 3. Elements and processes of career management	4. Emergence of change 5. Interfaces and process
Unique Attributes	
1. Attributes of reviews (proposes 5th dimension of context in addition to 4: food, service, ambience, and price) 2. Content (4) 3. Elements of virtuality 4. Errors associated with best practices 5. Features of cellphones/hard disk drives 6. Global integration and local responsiveness	7. Quantifying the bias 8. Tax collection and public spending 9. Themes of medication concerns 10. Topics covered in debate 11. Topic (structure, focus, newness) 12. Topics related to bridge management 13. Types of cultural conflict 14. Workgroup context

Note. Asterisks (*) indicate attributes that are potentially relevant to employment. The numbers in parentheses indicate the number of times that attribute appeared in the review.

We would also make the following observations on the attributes that have been measured in the literature:

1. Most of the research has focused on more macro topics, such as firm-level research on strategy, CEO and top management behavior, investor behavior, news reports, social media, and other organizational behavior topics (e.g., culture, climate, teams, leadership, communication), with fewer on micro topics such as individual employee behavior and human resources topics.

2. That may also explain the reliance on existing data sources, although the causal direction is uncertain. Did research using CATA use these sources because they are textual? Or did the availability of these data lead to the use of CATA because that was the only way to analyze such data?

3. Of those studies focusing on psychological variables, CATA has been used to measure states like attitudes, beliefs, etc., and traits like personality, knowledge, etc.

4. They also have been used to measure process variables, though this is largely in the inductive domain.

5. Measuring sentiment has been a dominant theme, perhaps due to the early development of that type of work dictionary. Research has measured sentiments toward things of all types such as opinions about products or places (hotel reviews; Hu, Chen, & Choi, 2017) and media favorability of a CEO or company (Love, Lim, & Bednar, 2017).

4.2 What Research Has Been Conducted in the Context of Employment?

Of particular relevance to the needs of the Air Force are studies related to measuring human attributes that might be useful for hiring (such as KSAOs relevant to entry-level hiring). We studied all the articles that appeared to measure constructs potentially relevant to KSAOs for staffing decisions (such as those indicated by asterisks in Table 4) as well as other staffing issues (e.g., recruitment and turnover). They are described briefly below in terms of how they used CATA to measure KSAOs and how a similar approach might be useful in the Air Force context. They are in alphabetical order.

Banks, Woznyj, Wesslen, Frear, Berka, Heggestad, and Gordon (2019) tested how strength of recruitment signals varied across domestic and international recruiting websites. They text mined and analyzed data from 162 websites across 21 countries using DICTION to develop themes. They found that signal strength on domestic sites was related to signal strength on international sites. The relevance to the Air Force is the possible use of text mining to understand and improve recruiting messages.

Buyl, Boone, and Wade (2019) analyzed CEO letters to shareholders and used LIWC to evaluate entrepreneurial orientation. While this measure was used simply to validate their proxy for CEO narcissism (not text analyzed), orientations are conceptualized as being stable and are therefore similar to traits, making their approach potentially useful in the context of employment. Two other papers measured entrepreneurial orientation using CATA and may also be useful here: Engelen, Neumann, and Schmidt (2016), and Short, Broberg, Cogliser, and Brigham (2010).

Campion, Campion, Campion, and Reider (2016) used text mining with LSA and NLP to measure several job-related skills (e.g., communication, critical thinking, people, leadership, managerial, and factual knowledge) based on accomplishment records in a hiring context. They validated these skill assessments against ratings by a panel of three trained assessors and achieved a correlation of .60 (the same as interrater reliability). This approach is highly relevant and one that we will likely try with the data from the Air Force.

Chen and Latham (2014) measured achievement by using the achievement dictionary in LIWC to analyze open-ended survey questions. Measuring achievement by using essays might be useful in a hiring context.

Connelly, Zweig, Webster, and Troukagos (2012) examined "knowledge hiding," or instances where newer organizational members are unwilling to share knowledge relevant to the new organization. They generated a list of 26 knowledge hiding behaviors, available in Table 1 in their article. Perhaps this research might be useful in gaining insight on reasons for turnover.

Felps (2009) analyzed transcripts from 11 focus groups using Atlas.ti and generated a list of job search behaviors. The frequency of use of these words by focus group members was correlated with job embeddedness, commitment, and satisfaction, which suggests that word choices of employees (that could be collected in a range of ways) may predict turnover intentions.

Kregel, Ogonek, and Matthies (2019) created a competency taxonomy of "lean manufacturing" professionals by using CATA with latent semantic analysis to content analyze job advertisements. They used CATA to extract the competency-specific words and then they used two independent raters to place them into a pre-defined framework. Such a technique could be used to extract competencies from applications in a hiring context.

Kwantes, Derbentseva, Lam, Vartanian, and Marmurek (2016) attempted to measure the Big Five (openness, conscientiousness, extraversion, agreeableness, and neuroticism) using CATA. They had undergraduates complete a personality test and also write five essays, one for each personality trait, in which they were asked to describe what they would do and how they would feel in each of five scenarios designed to invoke the creation of narrative relevant to the Big Five personality traits. They then used latent semantic analysis to analyze the text and correlated the scores with the personality test scores. They obtained significant correlations in the .19 to .32 range for three of the five traits (openness, extraversion, and neuroticism). Although the correlations are not that high, such an essay approach could be used in a hiring context to measure personality traits. An important question is whether the faking likelihood in a hiring context would be as great with an essay approach as it is with a direct survey approach. If not, this might be a viable alternative to measuring personality traits that would be more valid.

Liang et al. (2016) used CATA to measure positive-negative sentiment to assess why abusive supervisors are abusive. They did a sentiment analysis using LIWC of written descriptions of recalled interactions supervisors had with subordinates. A similar approach might be used to analyze descriptions of past behavior provided by officer job applicants to measure their potential for leadership.

Madera, Hebl, and Martin (2009) measured characteristics of candidates (communal versus agentic traits) in letters of recommendation using LIWC and their own dictionary. They also used human raters to support construct validity of LIWC by having them rate the letters directly and correlate the ratings with the LIWC scores, although they achieved only low correlations (in the teens). They also found a correlation between the LIWC scores and hireability ratings (of .29). The purpose of the study was to examine subgroup differences and show that female candidates were described in more communal terms while male candidates in more agentic terms. Nevertheless, measuring agency as a trait might be useful for hiring employees.

Malhotra, Reus, Zhu, and Roelofsen (2018) measured the extraversion of CEOs based on their spoken responses to questions using the LIWC psycholinguistic database. They also used two human raters as validity checks by rating the CEOs statements directly on extraversion items from the Big Five. Further, they correlated the answer length with the LIWC scores because longer answers are an indicator of extraversion. Measuring extraversion from candidate reactions could be useful in a hiring context, as is the finding that answer length is another simple measure of extraversion.

Martin (2016) examined how recent hires learned organizational values by using a "values" dictionary in LIWC to content analyze stories written by recent hires to illustrate the organization's values. The use in hiring might be to have potential candidates write descriptions of their past behavior or current life that illustrates their values (or write descriptions of how they embody the Air Force values).

McAlearney (2006) conducted a grounded theory study on leadership with 125 interviews and identified six high-level challenges to leadership development: industry lag, representativeness, professional conflict, time constraints, technical hurdles, and financial constraints. Measuring leadership traits is potentially important to hiring officer candidates.

Moore, Lee, Kim, and Cable (2017) used LIWC to measure self-verification striving (which is like authenticity) in employment interview transcripts to predict job offer likelihood. This is not only an example of the use of CATA in a hiring context, but it may suggest a new construct that could be useful given the pervasive problem of faking answers by candidates in interviews (i.e., impression management).

Onan and Korukoglu (2016) presented and tested an "ensemble approach" for feature selection in machine learning. Usually, we use a "filter approach" where features (different types of data or variables) are selected if they surpass some minimum threshold of usefulness. It is difficult to train a learning algorithm if there is too much variety of information and it degrades the model, points that the authors of this review have made to the Air Force. An ensemble approach aggregates features to make the process more efficient. One author of the current report reviewed a project conducted at another client (Amazon) using this approach to process applications, which showed some incremental value. Thus, this technique is worth noting, although special machine learning software may be needed to use it.

Owens and Hekman (2012) used Atlas.ti to conduct a content analysis to extract behaviors that symbolize humble leaders from interview transcripts. Then, two research assistants independently coded and verified the coding scheme. The authors stated that "leader humility involves leaders modeling to followers how to grow and produce positive organizational outcomes by leading followers to believe that their own developmental journeys and feelings of uncertainty are legitimate." They also said it is "leading from the ground" or "bottom-up leadership." Measuring leadership style, although perhaps not this exact version, could be relevant to selecting officers for the Air Force.

Ptriglieri (2015) studied identity reconstruction with your employer among executives at British Petroleum after the Gulf oil spill by conducting a content analysis of interview scripts to reveal 14 second-order themes and 5 aggregate dimensions. This paper illustrates how to use content analysis to identify psychological constructs relevant to the work setting. Further, identifying with the Air Force might itself be a useful construct to measure among candidates.

Reay, Goodrick, Waldorff, and Casebeer (2017) studied changes in professional role identity of physicians by conducting content analysis of interviews using Nvivo to discover 9 second-order codes and 4 aggregated codes. Two independent coders to helped develop keyword lists and two other independent coders to code the text. As with the Ptriglieri (2015) study, it showed how to develop second-order and then aggregated codes. This type of hierarchical coding is common. They measured a job-related construct that might be relevant to the Air Force and provided a good illustration of content analysis using text mining software.

Reincke and Ansari (2015) studied clock-time orientation using a content analysis using Nvivo with interview transcripts, observations, and archival documents. They found 92 descriptive codes, 29 first-order categories, 9 second-order themes, and 3 aggregate theoretical dimensions. Interesting points include starting out with descriptives, the large number of codes, and the concept of clock time, all of which seem relevant to the Air Force. Linear Clock-Oriented Time is linked to Western thought where time is considered to be an objective and quantifiable measure of motion, events, and actions. It is in contrast to process-oriented time, which is linked to Eastern thought where time is not autonomous or independent of events, processes, or phenomena, but characterized as non-linear, qualitatively determined, and endogenous to events and processes. The Air Force would presumably want employees with a linear clock orientation.

Ridge and Ingram (2017) assessed top management team modesty to predict returns and firm performance. While at the team rather than the individual level, their use of DICTION and Nvivo to analyze conference call transcripts demonstrates a way to capture a trait (modesty).

Rothausen, Henderson, Arnold, and Malshe (2017) studied identity to predict voluntary turnover by using Nvivo to content analyze interview transcripts. They found 18 second-order codes and 4 aggregate dimensions. The content categories might be of interest to the Air Force because they are potential predictors of turnover. Also, they used five interrelated procedures for qualitative research from Silverman and Marvasti (2008: pp. 257–270): refutability, constant comparison, comprehensive data treatment, deviant-case analysis, and respondent validation. They also did a back translation at the end by recoding each interview into the final categories.

Sajjadiani, Sojourner, Kammeyer-Mueller, and Mykerezi (2019) used machine learning to predict turnover among school teachers. This article was cited previously. It is the most recent study relevant to CATA that came out in the *Journal of Applied Psychology* in 2019. They did not analyze text in the form of essays, but instead analyzed applicant work history. Such a process could be used by the Air Force to predict the likely turnover of recruits at time of hire from their work history and other application information. Potentially, data already exists from past hires and turnover data to examine this issue. This is a modern-day version of using weighted application blanks and biographical data, which has a good history of validity for turnover (e.g., Reilly & Chao, 1982).

Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, Agrawal, ... and Ungar (2013) measured personality via text mining using LIWC "Differential language analysis (DLA)" and an "open-vocabulary approach" to mine Facebook messages. They showed that certain word usage related to personality traits as measured on personality tests. The authors claim that it is the largest study of language and personality. The multiple Rs between the words and the personality scores were .21 to .29, so the prediction was not great but, as noted above, this might be good if personality measured in this way is less fakable. They also found differences by age and gender, suggesting the importance of controlling for these in the use of language. They did not explore race differences.

Shantz and Latham (2009) measured need for achievement using LIWC to conduct a content analysis of handwritten stories. LIWC has 186 words related to achievement. The importance of this study is two-fold. First, achievement seems to be a general construct relevant to the Air Force. Second, this type of study demonstrates how we can analyze stories and draw out achievement orientations regardless of writing quality, as it is a concern of the United States Air Force Academy given the unequal distribution of educational quality across congressional districts.

Smither and Walker (2004) analyzed the favorableness and the trait versus behavioral nature of comments in 360 degree surveys. It is not clear that they used any software for the text analysis. It is included here because it did an analysis of text in context relevant to performance evaluation.

Sonenshein (2014) evaluated how organizational members engaged in "creative resourcing," which is essentially a technique focused on repurposing available materials to solve problems. Through coding a series of 55 interviews using Nvivo, he developed five aggregate themes: perceived resource endowment (capital resources, knowledge resources), resourcing identity (cultivating ownership schemas, cultivating creator schemas), regulating objects (withholding fixed objects, provisioning dynamic objects), creative resourcing (manipulating, recombining), and problem solving (original solutions, heterogeneous solutions). The construct of "creative resourcing" could be a potential skill in hiring.

Sorour, Goda, and Mine (2017) examined how student comments on learning predict performance using sentiment analysis with Mecab software, which uses LDA and Probabilistic Latent Semantic Analysis (pLSA). This is of interest to the Air Force not only because it predicts learning from student comments, including providing ideas for questions to ask students, but it illustrates how to use machine learning to predict outcomes. This complements the above studies that used simple content analysis software. The tradeoff is less construct meaning information and more prediction of important outcomes.

Speer (2018) text mined performance management narrative comments using R to conduct sentiment analysis. "The derived narrative scores were reliable across years, converged with traditional numerical ratings and explained incremental variance in future performance outcomes (performance ratings, involuntary turnover, promotions, and pay increases)" (p. 299). Although the article did not prove it directly, they also make the following useful point:

"… inclusion of narrative comments will result in (a) increases in total information and reliability, which is expected to occur across appraisal settings. In addition, in contexts where narratives are not explicitly linked to distributive outcomes (e.g., pay), narratives are likely to (b) exhibit a reduced amount of variance attributable to rater bias …. it is pertinent to note that motivation to distort will also likely exist. For example, just as a lower traditional rating could lead to an unpleasant confrontation with a subordinate, negative comments could also spur unpleasant interactions that promote avoidance motivations (Wang, Wong, & Kwong, 2010; Wong & Kwong, 2007). However, whereas this may occur for both mediums in likely equal probability, traditional narrative ratings are substantially more likely to be explicitly tied to distributive outcomes than narrative comments are, and therefore more likely to be affected by leniency bias" (pp. 304-306).

This article is useful to the Air Force because it may be the best recent example of using text mining for appraisals and the potential to gather both additional and candid information compared to ratings.

Wilhelmy et al. (2016) used Atlas.ti to conduct a content analysis of interview transcripts, observations, memos, and archival data to understand interviewer impression management (IM). They found 19 higher-level categories of IM tactics by interviewers. They trained coders to use Atlas.ti, then had them code independently. They used triangulation of transcripts of interviews, observations of interviews, and asking interviewers directly. After identifying categories, they linked them into a causal model (intentions to behaviors to outcomes). This is relevant to the Air Force because it is on interviews for hiring. Plus, the methodology seemed useful in several regards.

Of special note is that very limited research has been conducted on the topic of subgroup differences in text mining. Only four studies could be found that commented on subgroup differences. In a study of the extent to which managers of different genders delegate, Akinola, Martin, and Phillips (2018) found that women were more likely to associate being instructed to delegate with greater negative affect than men. In a study of using text mining of applicant accomplishments, Campion et al. (2016) found that text mining had no effect on adverse impact. They asserted that computer scoring against assessor scores will only produce adverse impact if there is already adverse impact in the assessor scores since the computer is simply modeling the assessors. In a study of why male entrepreneurs raise more funding, Kanze, Huang, Conley, and Higgins (2018) found that investors ask female entrepreneurs more prevention-focused questions and ask males entrepreneurs more promotion-oriented questions. Although not related to hiring, this type of study might be used to determine if military recruiters ask female candidates different questions than male candidates. Finally, as discussed previously, Madera et al. (2009) found that letter-of-reference writers used more communal terms for female applicants and more agentic terms for male candidates.

There were also several presentations at the 2019 Society for Industrial and Organizational Psychology conference on machine learning that included discussion of subgroup differences and the potential for adverse impact (e.g., Walmsley, 2019). Note that these conference presentations were not focused on CATA, but on ML using all types of applicant data. They did not address the issue above about creating models using human judge scores as the criteria that have adverse impact, but were instead focused more on developing or changing models that did not exhibit any subgroup differences. They stated that adverse impact is often observed when machine learning is applied to selection information (such as applications), and the solutions have been:

(a) Deny the computer from mining any information that might be illegal (e.g., race, gender, and age information). This might also include information that is associated with these protected categories (e.g., name of school attended). Campion et al. (2016) used this approach.

(b) Correlate the variables extracted by the computer with race and gender and eliminate those variables that show differences. However, this has a potential cost in terms of reducing validity. As commonly observed in employment testing, the items with the largest race differences are often the most valid items, so eliminating them reduces validity.

(c) Programming the algorithm to not have subgroup differences. This may potentially be tantamount to within-group norming, which is illegal based on the Civil Rights Act of 1991. This thought has also occurred to our contacts at the EEOC.

4.3 What Evidence Supports Effectiveness?

As best described in the reviews of the literature (Kobayashi et al., 2018a, 2018b; McKinney et al., 2016; Short et al., 2018), the evidence of effectiveness has been of two types: evaluation and validation. We described the types of validation evidence above under types of construct validity. The total number of studies on CATA in peer-reviewed journals and the amount of research using each type of construct validity evidence (Table 1) suggests that there is substantial evidence of validity. Because the studies have been on such a wide range of topics and used such a wide range of methodologies, statistical summaries (meta-analyses) are not possible. Perhaps such summaries will be possible in the future as CATA is used in specific contexts with consistent methodologies (e.g., for employment decisions used job performance as criteria).

Evaluation evidence includes a wide range of techniques focused on the quality of the computer model itself, rather than its usefulness. The primary types of evaluation evidence include the following:

1. Cross-validation. This is a central concern with CATA for several reasons. First, the research is often exploratory and inductive. Second, the computer programs are capable of capitalization on chance because they are designed to detect virtually any relationships in the data when being trained, which may include fitting the model to error variance to some extent. Third, the sheer volume of variables created during training means that some will be significant by chance. Fourth, when human judgment is used in the derivation of categories, even when

assisted by SMEs, there is a subjective component that may be different with other researchers or SMEs.

The approaches to cross-validation vary as well. They include both replicating on a new sample, splitting the sample and cross-validating in each subsample (including k-fold cross-validation, which is a re-sampling procedure to estimate the stability of the model) (e.g., Khan, Qamar, & Bashir, 2016a, 2016b; Yong, Tang, Cui, & Long, 2018), repeating subjective judgments with separate or additional SMEs (Gioia, Price, Hamilton, & Thomas, 2010, used SMEs and independent coders in addition to original coders; Ramarajan, Bezrukova, Jehn, & Euwema, 2011, used two sets of independent coders), and statistical techniques (e.g., shrinkage estimates). Of course, the use of large samples is another way to reduce the impact of capitalization on chance, though what constitutes "large" depends on the goals of the data. For example, grounded theory studies and other more rudimentary CATA requires smaller sample sizes of about 60 (Kreiner, Hollensbe, & Sheep, 2006, 2009) to 125 interviews (McClelland, Liang, & Barker, 2010). On the other hand, text mining requires significantly more power as algorithms can include more than 1,000 variables, which necessitates extremely large samples (Campion et al., 2016, used a sample of 42,000). A supplemental analysis in Campion et al, as well as two other articles cited by them, recommended a minimum of 500. The sample size needed depends on (a) the variation in the responses, (b) the amount of error in the data due to all sources (e.g., verbosity, misinterpretation, misspellings, unrelated text, unusual terms, etc.), (c) the reliability of the criterion predicted to develop the model (if used), (d) the desired standard errors (such as in the percent of responses in each category, or the mean on scores), and (e) the required level of model accuracy, and other factors. Minimum sample size will also depend on cross-validation plans.

2. Accuracy of classification. When using CATA to create categories, such as for content analysis, the accuracy of classification can be an important evaluation of the model. Such evidence may be relevant in several different contexts. For example, the interest might be in how well the computer can reproduce the categories created by humans or vice versa (e.g., Campion et al., 2016). It might be of interest how well the computer can classify observations into known categories (e.g., subgroups of people based on reasons for turnover; Sajjadiani et al., 2019). Or it might be of interest how well different computer models classify observations into the same categories.

3. Psychometric indices, such as reliability and error of measurement. As described earlier, psychometric evidence can be used to support the construct validity of the attributes identified. However, it also can serve the role of evaluating the quality of the resulting data from the perspective of traditional psychometric support. Although about 20% of studies included such evidence (Table 1), it would seem that this data would usually be relevant and easy to calculate in about any study. Even though these are variables created by computer algorithms, fundamental considerations of consistency, repeatability, and freedom from random error are still important.

4. Interpretably. Most researchers using CATA will usually rely to some extent on the logical appeal or ease of interpretation of the results to some extent. As noted earlier, SMEs will often be used to ensure the interpretability of the results. In academic research articles, the correspondence between the CATA results and the theory or hypotheses will be critical. In more exploratory research, the common-sense appeal may be relevant.

5. Operational considerations. Finally, and probably not the least relevant type of evidence bearing on evaluation, there are operational considerations creating costs and benefits. Example costs might include the time to build the model, programming skill necessary to do so, cost of the software, frequency that the computer algorithm must be updated, and so on. Example benefits might include saved labor cost by automating a human decision process, improved accuracy, reduced subjectivity and bias, increased speed, and others.

4.4 What can be Learned from other Disciplines that Use Similar Techniques?

As explained in the purpose section, the focus of this review is on the use of CATA in the management literature to measure human attributes, with a particular emphasis on construct validity. The vast literature on machine learning was not directly relevant because it did not focus on construct validity and human attributes, and usually not on text data. However, even though we did not review that literature comprehensively, we scanned the text mining literature that exists in other disciplines for lessons that might be relevant to the project goals. Potentially useful observations from that literature include the following:

1. The machine learning research outside management often includes both quantitative and text variables. So one lesson from that literature is that we should not limit ourselves just to text data and CATA applications of machine learning. Quantitative variables can be included in the same computer algorithms as the text-mined variables. Most often, the text-mined variables are converted into quantitative variables and included in the statistical model (like in a regression). In the case of assessment for staffing, other such data might include application information like years of work experience, years of education, grades, etc., or even test scores or other assessment information.

2. The focus in the literatures on prediction of outcomes over construct validity reveals that there are many other statistical models that might improve prediction. The management literature is unnecessarily narrow in its almost sole reliance on normal regression. Examples of other regression models available within SPSS include logistic regression (for dichotomous criteria), cox regression (for hazard models), autoregressive integrated moving-average model (for time series), discriminant analysis (for distinguishing between groups), Poisson regression (for low probability events), Gamma regression (for positive data ranges), and others. There are also alternatives to regression such as cluster analyses (k-means), classification models (decision trees, support vector machines, nearest neighbor, etc.), neural networks, and Bayesian networks. Some of these techniques have been used in the management literature, but they are not common. They may have advantages for improving the prediction of outcomes like job performance or turnover.

3. The range of outcomes predicted in those other literatures might give the Air Force ideas as to the broader applicability of these techniques. For example, considering employees as we consider customers in regards to satisfaction, loyalty, and representation of brand, research predicting customer behaviors from reviews (Bilro et al., 2019) may have insights for how to maintain loyalty with employees and strengthen their identification with the brand (e.g., Air Force reputation). Interestingly, a study on Amazon customer reviews showed that sentiment of reviews can be contagious and influence perceptions of how helpful a review is (Felbermayr & Nanopoulos, 2016). We may similarly see this emotional contagion occur and affect interpersonal relationships among military personnel. Further, researchers in management tend

to only look at job performance as a continuum, such as on a 5-point performance scale. Those in other literatures commonly predict specific categories (e.g., the likelihood someone will be rated a 5). Likewise, they would be more likely to try to predict sub-dimensions of job performance as opposed to just the overall composite like we do in management. Teasing apart the criterion in these ways can often lead to useful insights and predictions, as long as we properly attend to issues of capitalization on chance and spurious findings. The concept of "data mining" and its search of any and all possible relationships in the data may feel like a retreat to the "dustbowl empiricism" days in management in the past, but proper statistical treatment and more focus on construct validity should attenuate those concerns.

4. Researchers in other fields will often make adjustments to the distributions of the data, including both trimming and imputing data, and using assumed distributions. Management researchers tend to view the actual data collected as the best representation of the true phenomenon to be modeled or predicted, but these are usually convenience samples with known statistical limitations (e.g., restriction, missing data, skew, and others). As with relying on meta-analytic estimates of validity as opposed to performing a local validation (where statistical limitations reduce the chances of finding validity), sometimes it might be best to create a computer model based on the likely distribution rather than relying solely on available distributions with known problems. This is especially the case in the early stages of the research when the focus is on whether something "could" work as opposed to "will" work. This is much like conducting a lab study on a phenomenon to demonstrate its potential existence, and then following up with a field study to evaluate generalizability.

Data scientists will also use mathematical techniques to help solve problems with data distributions. As a specific example, one of the authors was asked to review a machine learning project for hiring at Amazon. The model used application information to predict recruiter decisions and subsequent job performance. The problem Amazon faced was that certain candidate characteristics would be almost completely missing from the data on hires because there are minimum qualifications and few would be hired who do not have those credentials. An example might be the possession of a college degree for hiring software engineers. They addressed the problem by using "inverse propensity weighting" wherein the model weights were adjusted to over-count the small sample of those hired without degrees.

5. Finally, the sources, styles, and amount of data are seemingly endless. As noted previously, this technique has broad applicability and includes data we have historically ignored as being useful for the purposes of traditional I-O topics. Further, the capacity of operating systems to process and manage huge amounts of data should reduce our apprehension about these data sources. For example, one study in our review text mined more than 1.7 million tweets and analyzed the sentiment of particular brands (e.g., Comcast) (Liu, Burns, & Hour, 2016). As an example more relevant to the Air Force, the U.S. Army recently tweeted out "How has serving impacted you?" With more than 12,000 responses and 9,600 retweets (which can sometimes include personal editorializing), text mining of this type of data may be useful for several purposes for the Air Force. Besides the more obvious assessment of recruitment efforts (how is the Air Force brand interpreted and how might that differ from the intended brand?), it may be useful in identifying issues related to retention and previously undiscovered motivations of potential military personnel. In the case of the Army, this tweet provoked an important and lively discussion about experiences in the Army—as well as other branches of the Armed Forces—that have the potential to positively inform operations within the Army. Of course, instances such as

these where there is entertainment value can sometimes yield false or faked responses, but information on social media remains potentially useful (Hartwell & Campion, 2019).

4.5 What are the Main Challenges in the Air Force's Intended Use?

Listed below are some of the challenges at each stage of using CATA for hiring assessments and other purposes.

1. Identifying viable applications. This will depend on many considerations, some of which are:

(a) Is textual data currently collected from the applicants? If not, could it be collected? As we have seen, no text data are collected from the enlisted candidates, but it is for officer candidates. A related consideration is whether the data are collected unproctored and, if so, will the candidate likely get help with the writing, and could this bias the scoring (e.g., writing skill)?

(b) Is the textual data collected likely to contain information on job-related constructs? If not, could the data collected be modified?

(c) Is there enough variance in the data to be useful for selection? For example, young candidates might not have enough work experience to use for text mining.

(d) Can a large enough sample of data be collected to create the CATA model? If not, could SMEs write illustrative text to train the model?

2. Deciding on software to use. As described elsewhere, the authors are inclined to use commercial software because it is more fully developed, user friendly, and documented, and technical support is more readily available. Also, some of the software used most commonly in the past literature is limited in its capabilities. However, we understand that the Air Force may have limitations as to the software they can use.

3. Learning the software. This will be a meaningful investment in staff time, even if commercial software is used. How much time will depend on many factors (e.g., computer skill, availability of a tutor, complexity of data, etc.)? We would estimate it will likely take several days to a week for an analyst to learn the basics, and then several weeks of additional time learning "on-the-job" as issues come up while applying the software over the next couple years.

4. Training the software. This is a challenge technically, of course, in that the training involves working with the computer model incrementally to improve it. With the SPSS software, training can be done in the form of revising the concepts or at the level of the categories of concepts. In either case, the training can take the form of combining, deleting, relabeling, identifying synonyms, and other adjustments. It may also include conducting studies to collect outcomes to train the algorithm against. This is also a question of how much is enough? Will training continue until some criterion level of accuracy is reached? In our previous research, the goal was to train the computer model until the correlation with a human assessor was the same as the correlation between assessors, which makes us ask whether training should continue until some asymptotic level is reached? Moreover, if no criterion is available to train against, how will the adequacy of the training be determined? Will it be the judgment of the programmer, or the appeal of the resulting model (e.g., categories created)?

5. Validating. This will depend on the approach taken:

(a) If criterion-validation is to be used, is there a criterion available or can one be developed to validate the model? Even if so, the usual criterion-related validity concerns will need to be addressed (e.g., restriction of range in the predictor, unreliability of the criterion, statistical power, etc.).

(b) If content validation is to be used, what will be the approach? Is a job analysis available? Are SMEs available? Will linkage analyses be conducted? Would the context meet the requirements for content validation in the Uniform Guidelines?

6. Updating the model to accommodate legal and social changes. As we have learned doing this work with another client, the trained models may become outdated due directly to legal changes according to equal employment regulations, but also due to social developments. Words can change or alter meaning over time. For example, the word "literally" is now well-understood to metaphorically mean "figuratively," and while the current generation was not the first to use it this way, they were the ones to make it popular (Merriam-Webster, n.d.; also see "Google" as a dictionary addition in 2006). Moreover, new words and phrases emerge from social movements (e.g., "black lives matter," "#metoo"). Modifying the algorithm requires human intervention to train the computer how adjusted or newly developed terms relate to existing terms. Of course, this may ultimately necessitate re-validation of the model should the changes be significant.

7. Working out operational details. As with any project, the "devil is in the details." Creating an application for operational use and putting it into practice is no exception and may be even more complicated these days due to the need for IT support and the interrelatedness of the IT systems. Issues include researcher support, programmer support, hardware requirements, flow of data and time, data cleaning, scoring, cut scores, data maintenance, security issues, other systems integration issues, etc.

8. Communicating to candidates. Although the use of CATA to help score textual data as part of the hiring process does not necessarily create additional needs to communicate with candidates, the usual communications may have to be adjusted. These will likely include communications posted on the website or in other recruiting documents as to the nature of the assessment used in the hiring process, any preparation advice, feedback on scores, retesting policies, responses to Freedom of Information Act (FOIA) requests, and so on. However, some candidates may still have mistrust in computer scoring that will influence their reaction. Perhaps communicating that the machine scoring is highly reliable and bias-free would mitigate their concerns, as fairness communications have been shown to improve reactions in other hiring contexts (e.g., Truxillo, Bauer, Campion, & Paronto, 2002).

4.6 What are the Recommendations for Using CATA by the Air Force?

Many recommendations have already been made explicitly or implicitly throughout the report, both for using CATA and many other topics. The current section summarizes our recommendations regarding the approach to use.

1. In terms of a bottom-line, we recommend that CATA be used as an approach to measuring KSAOs for employment decisions. The literature review suggests that CATA can and has been used to measure a range of constructs. Although relatively few have measured KSAOs directly related to employment decisions, the range of constructs measured in the literature suggests that

KSAOs could viably be measured. In fact, they might be measured more easily because they are more distinct and definable than many psychological constructs measured in the literature.

2. We recommend using more sophisticated methods than the CATA methods used in most of the management literature to date. They are probably not going to be sufficient alone for Air Force purposes. They are generally very simplistic and focus only on facilitating content analysis (e.g., Nvivo and Atlas). More sophisticated and better approaches are available. We specifically recommend using approaches to text mining that evaluate strings of words and relationships among words like LSA and NLP, as opposed to the more common simple single word and phrase-based approaches.

3. We recommend using approaches that allow training. Approaches vary in terms of whether and how easily they can be trained. The literature will often use the terms "supervised versus unsupervised" to make this distinction. The Air Force will want to use approaches that allow training because it can greatly improve interpretation and prediction. Another value of training is that it helps ensure that the researcher understands what the computer is scoring instead of blindly accepting the "black box." This is much the same reason researchers must learn how to calculate statistics by hand in graduate school, even though they will use computers in the future. Parenthetically, the layperson interpretation of "machine learning" is that the computer teaches itself. However, in the current state of development of the field, this refers more to the fact that the computer can fit a model to the data (like regression has done for years), rather than some continuous self-teaching from moment to moment. Researchers still have to train the computer in many ways even with today's modern software.

4. That said, we also recommend not ignoring the use of word dictionaries. They offer at least two advantages. First, many dictionaries have already been developed and validated on a range of constructs, and they are inexpensive to purchase (e.g., DICTION and LIWC). Second, it is fairly easy to develop one's own dictionary on constructs of interest. Although data dictionaries are perhaps the most simplistic approach to text mining, they can be developed in advance based on theory and do not rely on a corpus to text mine necessarily as the first step like the more sophisticated approaches. We note that SPSS contains a set of dictionaries and it allows the user to create their own. However, it does not contain all the dictionaries in LIWC, so LIWC might also be used.

5. Related to that, we recommend considering the various ways to strategically select or create corpuses (corpora) for developing a CATA model. For example, documents could be selected that are likely to be enriched with words relevant to constructs of interest in order to identify words for dictionaries (e.g., documents describing leadership might be used to identify leadership-related word descriptions). Similarly, illustrative text samples created by SMEs can be text mined to measure constructs better than using actual examples from subjects. Creating an algorithm for scoring constructs does not always start with actual examples from the future intended corpora.

6. We also recommend considering sentiment analysis as it might be appropriate. It could conceivably be used anytime it is necessary to distinguish positiveness/negativeness of comments. This has obvious applications when analyzing survey responses where the tone of the comments is important along with the content. It could potentially help resolve perhaps the most central problems in performance evaluation—leniency and skew in the ratings due to unwillingness on the part of managers to give candid feedback. Text-based appraisal

information that is somewhat less susceptible to this issue is the narrative comments used to explain the ratings (Speer, 2018). These comments are less susceptible because, even though they will be consistent with the ratings, they can reveal the strength of the job performance through what is sometimes called “faint praise.” Sentiment-based CATA might be able to measure these nuances more objectively. Similarly, employers want candidates with a “positive attitude,” but this is vulnerable to faking. Using sentiment analysis to assess attitudes from written materials may make response bias more difficult because it is less direct (less obvious to candidates) and it will be more difficult to know which word choices are scored. SPSS may be able to accommodate sentiment analysis, but the other software products that have emphasized this type of analysis (especially LIWC) might be needed.

REFERENCES

Abulaish, M., Jahiruddin, & Bhardwaj, A. (2019). OMCR: An opinion-based multi-criteria ranking approach. *Journal of Intelligent and Fuzzy Systems*. (Advanced online publication.)

Akinola, M., Martin, A. E., & Phillips, K. W. (2018). To delegate or not to delegate: Gender differences in affective associations and behavioral responses to delegation. *Academy of Management Journal, 61(4)*, 1467-1491.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Ashforth, B. E., Kreiner, G. E., & Fugate, M. (2000). All in a day's work: Boundaries and micro role transitions. *Academy of Management Review, 25*(3), 472 – 491.

Banks, G. C., Woznyj, H. M., Wesslen, R. S., Frear, K. A., Berka, G., Heggestad, E. D., & Gordon, H. L. (2019). Strategic Recruitment across Borders: An Investigation of Multinational Enterprises. *Journal of Management, 45*(2), 476-509.

Banks, G. C., Woznyj, H. M., Wesslen, R. S., & Ross, R. L. (2018). A review of best practice recommendations for text analysis in R (and a user-friendly app). *Journal of Business and Psychology, 33(4),* 445-459.

Barlow, M. A., Verhaal, J. C., & Hoskins, J. D. (2018). Guilty by association: Product-level category stigma and audience expectations in the US craft beer industry. *Journal of Management, 44* (7), 2934-2960.

Beckman, C. M., & Stanko, T. L. (2019). It takes three: Relational boundary work, Resilience, and commitment among navy couples. *Academy of Management Journal*. (Advanced online publication.)

Bednar, M. K. (2012). Watchdog or lapdog? A behavioral view of the media as a corporate governance mechanism. *Academy of Management Journal, 55(1)*, 131-150.

Belderbos, R., Grabowska, M., Leten, B., Kelchtermans, S., & Ugur, N. (2017). On the use of computer-aided text analysis in international business research. *Global Strategy Journal, 7*(3), 312-331.

Bilro, R. G., Loureiro, S. M. C., & Guerreiro, J. (2019). Exploring online customer engagement with hospitality products and its relationship with involvement, emotional states, experience and brand advocacy. *Journal of Hospitality Marketing & Management, 28(2)*, 147-171.

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

Bligh, M. C., Kohles, J. C., & Meindl, J. R. (2004). Charisma under crisis: Presidential leadership, rhetoric, and media responses before and after the September 11th terrorist attacks. *The Leadership Quarterly, 15(2),* 211-239.

Brett, J. M., Olekalns, M., Friedman, R., Goates, N., Anderson, C., & Lisco, C. C. (2007). Sticks and stones: Language, face, and online dispute resolution. *Academy of Management Journal, 50(1),* 85-99.

Brownlee, J. (2013, November 17). What is machine learning? Retrieved from https://machinelearningmastery.com/what-is-machine-learning/

Buyl, T., Boone, C., & Wade, J. B. (2019). CEO narcissism, risk-taking, and resilience: An empirical analysis in US commercial banks. *Journal of Management, 45(4),* 1372-1400.

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology, 101(7),* 958 - 975.

Carley, K. M. (1997). Extracting team mental models through textual analysis. *Journal of Organizational Behavior, 18*, 533-558.

Caza, B. B., Moss, S., & Vough, H. (2018). From synchronizing to harmonizing: The process of authenticating multiple work identities. *Administrative Science Quarterly, 63(4),* 703 – 745.

Chatman, J. A., Caldwell, D. F., O'Reilly, C. A., & Doerr, B. (2014). Parsing organizational culture: How the norm for adaptability influences the relationship between culture consensus and financial performance in high-technology firms. *Journal of Organizational Behavior, 35(6),* 785-808.

Chen, X., & Latham, G. P. (2014). The effect of priming learning vs. performance goals on a complex task. *Organizational Behavior and Human Decision Processes, 125(2),* 88-97.

Chuang, A., Hsu, R. S., Wang, A. C., & Judge, T. A. (2015). Does west “fit” with east? In search of a Chinese model of person–environment fit. *Academy of Management Journal, 58(2),* 480-510.

Colquitt, J. A., Long, D. M., Rodell, J. B., & Halvorsen-Ganepola, M. D. (2015). Adding the “in” to justice: A qualitative and quantitative investigation of the differential effects of justice rule adherence and violation. *Journal of Applied Psychology, 100(2)*, 278 - 297.

Connelly, C. E., Zweig, D., Webster, J., & Trougakos, J. P. (2012). Knowledge hiding in organizations. *Journal of Organizational Behavior, 33(1),* 64-88.

Connolly, J. J., & Viswesvaran, C. (2000). The role of affectivity in job satisfaction: A meta-analysis. *Personality and Individual Differences*, *29(2),* 265-281.

Crayne, M. P., & Hunter, S. T. (2018). Historiometry in organizational science: Renewed attention for an established research method. *Organizational Research Methods, 21(1)*, 6-29.

Creed, W. D., DeJordy, R., & Lok, J. (2010). Being the change: Resolving institutional contradiction through identity work. *Academy of Management Journal, 53(6),* 1336-1364.

DICTION. (2019). Retrieved from https://www.dictionsoftware.com/diction-overview/

Dormann, C., & Zapf, D. (2001). Job satisfaction: A meta-analysis of stabilities. *Journal of Organizational Behavior*, *22(5),* 483-504.

Drisko, J. W., & Maschi, T. (2016). *Content analysis*. Oxford, UK: Oxford University Press.

Edmondson, A. C., & McManus, S. E. (2007). Methodological fit in management field research. *Academy of Management Review, 32(4),* 1246-1264.

Engelen, A., Neumann, C., & Schmidt, S. (2016). Should entrepreneurially oriented firms have narcissistic CEOs? *Journal of Management, 42(3),* 698-721.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Adoption by four agencies of uniform guidelines on employee selection procedures. *Federal Register, 43,* 38290-38315.

Felbermayr, A., & Nanopoulos, A. (2016). The role of emotions for the perceived usefulness in online customer reviews. *Journal of Interactive Marketing, 36,* 60-76.

Felps, W., Mitchell, T. R., Hekman, D. R., Lee, T. W., Holtom, B. C., & Harman, W. S. (2009). Turnover contagion: How coworkers' job embeddedness and job search behaviors influence quitting. *Academy of Management Journal, 52*(3), 545-561.

Gangloff, K. A., Connelly, B. L., & Shook, C. L. (2016). Of scapegoats and signals: Investor reactions to CEO succession in the aftermath of wrongdoing. *Journal of Management, 42(6),* 1614-1634.

Gelfand, M. J., Severance, L., Lee, T., Bruss, C. B., Lun, J., Abdel-Latif, A. H., ... & Moustafa Ahmed, S. (2015). Culture and getting to yes: The linguistic signature of creative agreements in the United States and Egypt. *Journal of Organizational Behavior, 36*(7), 967-989.

Gephart, R. P. (2004). Qualitative research in the Academy of Management Journal. *Academy of Management Journal, 47(4)*, 454 – 462.

Glaser, B., & Strauss, A. (1967). Grounded theory: The discovery of grounded theory. *Sociology the Journal of the British Sociological Association*, *12*(*1*), 27-49.

Gioia, D. A., Price, K. N., Hamilton, A. L., & Thomas, J. B. (2010). Forging an identity: An insider-outsider study of processes involved in the formation of organizational identity. *Administrative Science Quarterly, 55(1),* 1-46.

Godnov, U., & Redek, T. (2018). Good food, clean rooms and friendly staff: Implications of user-generated content for Slovenian skiing, sea and spa hotels' management. *Management: Journal of Contemporary Management Issues, 23(1)*, 29-57.

Gomulya, D., Wong, E. M., Ormiston, M. E., & Boeker, W. (2017). The role of facial appearance on CEO selection after firm misconduct. *Journal of Applied Psychology, 102(4)*, 617 – 635.

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.

Hájek, P. (2018). Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns. *Neural Computing and Applications, 29(7)*, 343-358.

Hannigan, T., Haans, R. F. J., Vakili, K., Tchalian, H., Glaser, V., Wang, M., Kaplan, S., & Jennings, P. D. (2019). Topic modeling in management research: Rendering new theory from textual data. *Academy of Management Annals.* (Advanced online publication.)

Hartwell, C. J., & Campion, M. A. (2019). *Assessing online identities: Recruiter perceptions of applicant social media content.* Manuscript under review.

Hirsh, J. B., & Peterson, J. B. (2009). Personality and language use in self-narratives. *Journal of Research in Personality, 43*, 524 – 527.

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15(9),* 1277-1288.

Hu, Y. H., Chen, Y. L., & Chou, H. L. (2017). Opinion mining from online hotel reviews–A text summarization approach. *Information Processing & Management, 53(2),* 436-449.

Janasik, N., Honkela, T., & Bruun, H. (2009). Text mining in qualitative research: Application of an unsupervised learning method. *Organizational Research Methods, 12(3),* 436-460.

Kabanoff, B., & Holt, J. (1996). Changes in the espoused values of Australian organizations 1986—1990. *Journal of Organizational Behavior, 17(3),* 201-219.

Kabanoff, B., Waldersee, R., & Cohen, M. (1995). Espoused values and organizational change themes. *Academy of Management Journal, 38(4),* 1075-1104.

Kakol, M., Nielek, R., & Wierzbicki, A. (2017). Understanding and predicting web content credibility using the Content Credibility Corpus. *Information Processing & Management, 53(5),* 1043-1061.

Kanze, D., Huang, L., Conley, M. A., & Higgins, E. T. (2018). We ask men to win and women not to lose: Closing the gender gap in startup funding. *Academy of Management Journal, 61(2),* 586-614.

Khan, F. H., Qamar, U., & Bashir, S. (2016). Multi-objective model selection (MOMS)-based semi-supervised framework for sentiment analysis. *Cognitive Computation, 8(4),* 614-628.

Khan, F. H., Qamar, U., & Bashir, S. (2016). Senti-CS: Building a lexical resource for sentiment analysis using subjective feature selection and normalized Chi-Square-based feature weight generation. *Expert Systems, 33(5),* 489-500.

King, E. B., Shapiro, J. R., Hebl, M. R., Singletary, S. L., & Turner, S. (2006). The stigma of obesity in customer service: A mechanism for remediation and bottom-line consequences of interpersonal discrimination. *Journal of Applied Psychology, 91(3),* 579 - 593.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihok, G., & Den Hartog, D. N. (2018a). Text classification for organizational researchers: A tutorial. *Organizational Research Methods, 21(3),* 766-799.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihok, G., & Den Hartog, D. N. (2018b). Text mining in organizational research. *Organizational Research Methods, 21(3)*, 733-765.

König, A., Mammen, J., Luger, J., Fehn, A., & Enders, A. (2018). Silver bullet or ricochet? CEOs' use of metaphorical communication and infomediaries' evaluations. *Academy of Management Journal, 61(4)*, 1196-1230.

Kregel, I., Ogonek, N., & Matthies, B. (2019). Competency profiles for lean professionals–an international perspective. *International Journal of Productivity and Performance Management, 68(2),* 423-446.

Kreiner, G. E., Hollensbe, E. C., & Sheep, M. L. (2006). Where is the "me" among the "we"? Identity work and the search for optimal balance. *Academy of Management Journal, 49*(*5*), 1031-1057.

Kreiner, G. E., Hollensbe, E. C., & Sheep, M. L. (2009). Balancing borders and bridges: Negotiating the work-home interface via boundary work tactics. *Academy of Management Journal, 52(4)*, 704-730.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage.

Kwantes, P. J., Derbentseva, N., Lam, Q., Vartanian, O., & Marmurek, H. H. (2016). Assessing the Big Five personality traits with latent semantic analysis. *Personality and Individual Differences, 102*, 229-233.

Lanaj, K., Foulk, T. A., & Erez, A. (2019). Energizing leaders via self-reflection: A within-person field experiment. *Journal of Applied Psychology, 104(1)*, 1-18.

Liang, L. H., Lian, H., Brown, D. J., Ferris, D. L., Hanig, S., & Keeping, L. M. (2016). Why are abusive supervisors abusive? A dual-system self-control model. *Academy of Management Journal, 59(4),* 1385-1406.

Lingo, E. L., & O'Mahony, S. (2010). Nexus work: Brokerage on creative projects. *Administrative Science Quarterly, 55(1)*, 47-81.

Locke, K. (2001). *Grounded theory in management research*. Thousand Oaks CA: Sage.

Love, E. G., Lim, J., & Bednar, M. K. (2017). The face of the firm: The influence of CEOs on corporate reputation. *Academy of Management Journal, 60*(*4*), 1462-1481.

Luciano, M. M., Mathieu, J. E., Park, S., & Tannenbaum, S. I. (2018). A fitting approach to construct and measurement alignment: The role of big data in advancing dynamic theories. *Organizational Research Methods, 21(3),* 592-632.

Machine Learning (2019). Retrieved from https://en.wikipedia.org/wiki/Machine_learning

Madera, J. M., Hebl, M. R., & Martin, R. C. (2009). Gender and letters of recommendation for academia: agentic and communal differences. *Journal of Applied Psychology, 94(6),* 1591.

Malhotra, S., Reus, T. H., Zhu, P., & Roelofsen, E. M. (2018). The acquisitive nature of extraverted CEOs. *Administrative Science Quarterly, 63(2),* 370-408.

March, J. G. 1991. Exploration and exploitation in organizational learning. *Organization Science, 2*, 71-87.

Martin, S. R. (2016). Stories about values and valuable stories: A field experiment of the power of narratives to shape newcomers' actions. *Academy of Management Journal, 59(5),* 1707-1724.

McAlearney, A. S. (2006). Leadership development in healthcare: a qualitative study. Journal of Organizational Behavior: *The International Journal of Industrial, Occupational and Organizational Psychology and Behavior, 27(7),* 967-982.

McClelland, P. L., Liang, X., & Barker III, V. L. (2010). CEO commitment to the status quo: Replication and extension using content analysis. *Journal of Management, 36(5),* 1251-1277.

McKenny, A. F., Short, J. C., & Payne, G. T. (2013). Using computer-aided text analysis to elevate constructs: An illustration using psychological capital. *Organizational Research Methods, 16(1),* 152-184.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 5*, 1093 – 1113.

Menon, A., Choi, J., & Tabakovic, H. (2018, July). What you say your strategy is and why it matters: Natural language processing of unstructured text. In *Academy of Management Proceedings* (Vol. 2018, No. 1, p. 18319). Briarcliff Manor, NY 10510: Academy of Management.

Merriam-Webster (n.d.). Retrieved from https://www.merriam-webster.com/words-at-play/misuse-of-literally

Moore, C., Lee, S. Y., Kim, K., & Cable, D. M. (2017). The advantage of being oneself: The role of applicant self-verification in organizational hiring decisions. *Journal of Applied Psychology, 102(11),* 1493-1513.

Morris, R. (1994). Computerized content analysis in management research: A demonstration of advantages and limitations. *Journal of Management, 20*, 903 – 931.

Mossholder, K. W., Settoon, R. P., Harris, S. G., & Armenakis, A. A. (1995). Measuring emotion in open-ended survey responses: An application of textual data analysis. *Journal of Management, 21(2),* 335-355.

Nadkarni, S., & Chen, J. (2014). Bridging yesterday, today, and tomorrow: CEO temporal focus, environmental dynamism, and rate of new product introduction. *Academy of Management Journal, 57(6),* 1810-1833.

Ng, T. W., Yam, K. C., & Aguinis, H. (2019). Employee perceptions of corporate social responsibility: Effects on pride, embeddedness, and turnover. *Personnel Psychology*, *72(1)*, 107-137.

Onan, A., & Korukoğlu, S. (2017). A feature selection model based on genetic rank aggregation for text sentiment classification. *Journal of Information Science, 43(1)*, 25-38.

Owens, B. P., & Hekman, D. R. (2012). Modeling how to grow: An inductive examination of humble leader behaviors, contingencies, and outcomes. *Academy of Management Journal, 55(4),* 787-818.

Pandey, S., & Pandey, S. K. (2017). Applying natural language processing capabilities in computerized textual analysis to measure organizational culture. *Organizational Research Methods, 22(3),* 765-797.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric properties of LIWC2007. Retrieved from http://www.liwc.net/LIWC2007LanguageManual.pdf

Pratt, M. G. (2009). For the lack of boilerplate: Tips on writing up (and reviewing) qualitative research. *Academy of Management Journal, 52(5),* 856 – 862.

Petriglieri, J. L. (2015). Co-creating relationship repair: Pathways to reconstructing destabilized organizational identification. *Administrative Science Quarterly, 60(3),* 518-557.

Quigley, T. J., Hubbard, T. D., Ward, A., & Graffin, S. D. (2019). Unintended consequences: Information releases and CEO stock option grants. *Academy of Management Journal*. (Advanced online publication.)

Ramarajan, L., Bezrukova, K., Jehn, K. A., & Euwema, M. (2011). From the outside in: The negative spillover effects of boundary spanners' relations with members of other organizations. *Journal of Organizational Behavior, 32(6),* 886-905.

Reay, T., Goodrick, E., Waldorff, S. B., & Casebeer, A. (2017). Getting leopards to change their spots: Co-creating a new professional role identity. *Academy of Management Journal, 60(3),* 1043-1070.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35(1),* 1-62.

Reinecke, J., & Ansari, S. (2015). When times collide: Temporal brokerage at the intersection of markets and developments. *Academy of Management Journal, 58(2)*, 618-648.

Reinert, M. (1990). Alceste, une me ´thodologie d'analyse des donne ´es textuelles et une application: Aure ´lia de Ge ´rard de Nerval. *Bulletin de Me ´thodologie Sociologique 26,* 24–54.

Ridge, J. W., & Ingram, A. (2017). Modesty in the top management team: Investor reaction and performance implications. *Journal of Management, 43(4),* 1283-1306.

Rothausen, T. J., Henderson, K. E., Arnold, J. K., & Malshe, A. (2017). Should I stay or should I go? Identity and well-being in sensemaking about retention and turnover. *Journal of Management, 43(7),* 2357-2385.

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology*. (Advanced online publication.)

SAS. (2019). Natural language processing: What it is and why it matters. Retrieved from https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html

Sbalchiero, S., & Tuzzi, A. (2016). Scientists' spirituality in scientists' words. Assessing and enriching the results of a qualitative analysis of in-depth interviews by means of quantitative approaches. *Quality & Quantity, 50(3), 1*333-1348.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kisinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one, 8(9),* e73791.

Shantz, A., & Latham, G. P. (2009). An exploratory field experiment of the effect of subconscious and conscious goals on employee performance. *Organizational Behavior and Human Decision Processes, 109(1),* 9-17.

Short, J. C., Broberg, J. C., Cogliser, C. C., & Brigham, K. H. (2010). Construct validation using computer-aided text analysis (CATA) an illustration using entrepreneurial orientation. *Organizational Research Methods, 13(2),* 320-347.

Short, J. C., McKenny, A. F., & Reid, S. W. (2018). More than words? Computer-aided text analysis in organizational behavior and psychology research. *Annual Review of Organizational Psychology and Organizational Behavior, 5,* 415-435.

Shortt, H. L., & Warren, S. K. (2019). Grounded visual pattern analysis: Photographs in organizational field studies. *Organizational Research Methods, 22(2),* 539-563.

Silverman, D., & Marvasti, A. (2008). *Doing qualitative research: A comprehensive guide*. Newbury Park, NJ: Sage.

Slutskaya, N., Game, A. M., & Simpson, R. C. (2018). Better together: Examining the role of collaborative ethnographic documentary in organizational research. *Organizational Research Methods, 21(2),* 341-365.

Smither, J. W., & Walker, A. G. (2004). Are the characteristics of narrative comments related to improvement in multirater feedback ratings over time?. J*ournal of Applied Psychology, 89(3),* 575.

Society for Industrial and Organizational Psychology. (2018). *Principles for the validation and use of personnel selection procedures* (5th edition). Bowling Green, OH: Author.

Sonenshein, S. (2014). How organizations foster the creative use of resources. A*cademy of Management Journal, 57(3),* 814-848.

Sorour, S. E., Goda, K., & Mine, T. (2017). Comment data mining to estimate student performance considering consecutive lessons. *Journal of Educational Technology & Society, 20(1),* 73.

Speer, A. B. (2018). Quantifying with words: An investigation of the validity of narrative-derived performance scores. *Personnel Psychology, 71(3),* 299-333.

Strauss, A. & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, NJ: Sage.

Truxillo, D. M., Bauer, T. N., Campion, M. A., & Paronto, M. E. (2002). Selection fairness information and applicant reactions: A longitudinal field study. *Journal of Applied Psychology, 87,* 1020-1031.

Vagnani, G. (2015). Exploration and long-run organizational performance: the moderating role of technological interdependence. *Journal of Management, 41*(*6*)*,* 1651-1676.

Walmsley, P. T. (2019). *Using machine learning and deep learning in hiring: Ethical, legal, and practical concerns.* Panel discussion presented at the Society for Industrial and Organizational Psychology, National Harbor, Maryland.

Wang, X. M., Wong, K. F. E., & Kwong, J. Y. Y. (2010). The roles of rater goals and ratee performance levels in the distortion of performance ratings. *Journal of Applied Psychology, 95,* 546.

Wiedner, R., Barrett, M., & Oborn, E. (2017). The emergence of change in unexpected places: Resourcing across organizational practices in strategic change. *Academy of Management Journal, 60(3),* 823-854.

Wilhelmy, A., Kleinmann, M., König, C. J., Melchers, K. G., & Truxillo, D. M. (2016). How and why do interviewers try to make impressions on applicants? A qualitative study. *Journal of Applied Psychology, 101(3),* 313 - 332.

Wilson, K. S., DeRue, D. S., Matta, F. K., Howe, M., & Conlon, D. E. (2016). Personality similarity in negotiations: Testing the dyadic effects of similarity in interpersonal traits and the use of emotional displays on negotiation outcomes. *Journal of Applied Psychology, 101(10),* 1405 - 1421.

Wong, K. F. E., & Kwong, J. Y. Y. (2007). Effects of rater goals on rating patterns: Evidence from an experimental field study. *Journal of Applied Psychology, 92,* 577–585.

Yong, S. H. I., Tang, Y. R., Cui, L. X., & Wen, L. O. N. G. (2018). A text mining based study of investor sentiment and its influence on stock returns. *Economic Computation & Economic Cybernetics Studies & Research, 52(1)*, 183-199.

Zuzul, T. (2018). “Matter battles:” Boundary objects and the failure of collaboration in two smart cities. *Academy of Management Journal, 62(3),* 739-764.

APPENDIX A: Literature Review List of Articles

Abrahamson, E., & Eisenman, M. (2008). Employee-management techniques: transient fads or trending fashions? *Administrative Science Quarterly, 53(4)*, 719-744.

Abrahamson, E., & Fairchild, G. (1999). Management fashion: Lifecycles, triggers, and collective learning processes. *Administrative Science Quarterly, 44(4),* 708-740.

Abrahamson, E., & Hambrick, D. C. (1997). Attentional homogeneity in industries: The effect of discretion. *Journal of Organizational Behavior, 18(S1)*, 513-532.

Abulaish, M., Jahiruddin, & Bhardwaj, A. (2019). OMCR: An opinion-based multi-criteria ranking approach. *Journal of Intelligent and Fuzzy Systems,* 1 - 15.

Aghababaei, S., & Makrehchi, M. (2016, October). Mining social media content for crime prediction. In 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI) (pp. 526-531). IEEE.

Akinola, M., Martin, A. E., & Phillips, K. W. (2018). To delegate or not to delegate: Gender differences in affective associations and behavioral responses to delegation. *Academy of Management Journal, 61(4),* 1467-1491.

Anand, N., & Watson, M. R. (2004). Tournament rituals in the evolution of fields: The case of the Grammy Awards. *Academy of Management journal, 47(1),* 59-80.

Antons, D., Joshi, A. M., & Salge, T. O. (2018). Content, contribution, and knowledge consumption: Uncovering hidden topic structure and rhetorical signals in scientific texts. *Journal of Management, 45(7),* 0149206318774619.

Banks, G. C., Woznyj, H. M., Wesslen, R. S., Frear, K. A., Berka, G., Heggestad, E. D., & Gordon, H. L. (2019). Strategic recruitment across borders: An investigation of multinational enterprises. *Journal of Management, 45(2),* 476-509.

Banks, G. C., Woznyj, H. M., Wesslen, R. S., & Ross, R. L. (2018). A review of best practice recommendations for text analysis in R (and a user-friendly app). *Journal of Business and Psychology, 33(4),* 445-459.

Barclay, L. J., & Skarlicki, D. P. (2009). Healing the wounds of organizational injustice: Examining the benefits of expressive writing. *Journal of Applied Psychology, 94(2),* 511 - 523.

Barlow, M. A., Verhaal, J. C., & Hoskins, J. D. (2018). Guilty by association: Product-level category stigma and audience expectations in the US craft beer industry. *Journal of Management, 44(7),* 2934-2960.

Basu, S., & Savani, K. (2017). Choosing one at a time? Presenting options simultaneously helps people make more optimal decisions than presenting options sequentially. *Organizational Behavior and Human Decision Processes, 139,* 76-91.

Beckman, C. M., & Stanko, T. L. (in press.) It takes three: Relational boundary work, resilience, and commitment among navy couples. *Academy of Management Journal.*

Bednar, M. K. (2012). Watchdog or lapdog? A behavioral view of the media as a corporate governance mechanism. *Academy of Management Journal, 55(1),* 131-150.

Belderbos, R., Grabowska, M., Leten, B., Kelchtermans, S., & Ugur, N. (2017). On the use of computer-aided text analysis in international business research. *Global Strategy Journal, 7(3),* 312-331.

Ben-Menahem, S. M., Von Krogh, G., Erden, Z., & Schneider, A. (2016). Coordinating knowledge creation in multidisciplinary teams: Evidence from early-stage drug discovery. *Academy of Management Journal, 59(4),* 1308-1338.

Bezrukova, K., Spell, C. S., Caldwell, D., & Burger, J. M. (2016). A multilevel perspective on faultlines: Differentiating the effects between group-and organizational-level faultlines. *Journal of Applied Psychology, 101(1),* 86 - 107.

Bezrukova, K., Thatcher, S., Jehn, K. A., & Spell, C. S. (2012). The effects of alignments: examining group faultlines, organizational cultures, and performance. *Journal of Applied Psychology, 97(1),* 77 - 92.

Bhattacharya, S., Yang, C., Srinivasan, P., & Boynton, B. (2016). Perceptions of presidential candidates' personalities in twitter. *Journal of the Association for Information Science and Technology, 67(2),* 249-267.

Bilro, R. G., Loureiro, S. M. C., & Guerreiro, J. (2019). Exploring online customer engagement with hospitality products and its relationship with involvement, emotional states, experience and brand advocacy. *Journal of Hospitality Marketing & Management, 28(2),* 147-171.

Bingham, C. B., & Kahl, S. J. (2013). The process of schema emergence: Assimilation, deconstruction, unitization and the plurality of analogies. *Academy of Management Journal, 56(1),* 14-34.

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

Bligh, M. C., Kohles, J. C., & Meindl, J. R. (2004). Charisma under crisis: Presidential leadership, rhetoric, and media responses before and after the September 11th terrorist attacks. *The Leadership Quarterly, 15(2),* 211-239.

Brett, J. M., Olekalns, M., Friedman, R., Goates, N., Anderson, C., & Lisco, C. C. (2007). Sticks and stones: Language, face, and online dispute resolution. *Academy of Management Journal, 50(1),* 85-99.

Bruns, H. C. (2013). Working alone together: Coordination in collaboration across domains of expertise. *Academy of Management journal, 56(1),* 62-83.

Buyl, T., Boone, C., & Wade, J. B. (2019). CEO narcissism, risk-taking, and resilience: An empirical analysis in US commercial banks. *Journal of Management, 45(4),* 1372-1400.

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology, 101(7),* 958 - 975.

Carley, K. M. (1997). Extracting team mental models through textual analysis. *Journal of Organizational Behavior, 18(S1),* 533-558.

Caza, B. B., Moss, S., & Vough, H. (2018). From synchronizing to harmonizing: The process of authenticating multiple work identities. *Administrative Science Quarterly, 63(4),* 703-745.

Chatman, J. A., Caldwell, D. F., O'Reilly, C. A., & Doerr, B. (2014). Parsing organizational culture: How the norm for adaptability influences the relationship between culture consensus and financial performance in high-technology firms. *Journal of Organizational Behavior, 35(6),* 785-808.

Chen, X., & Latham, G. P. (2014). The effect of priming learning vs. performance goals on a complex task. *Organizational Behavior and Human Decision Processes, 125(2),* 88-97.

Chuang, A., Hsu, R. S., Wang, A. C., & Judge, T. A. (2015). Does West "fit" with East? In search of a Chinese model of person–environment fit. *Academy of Management Journal, 58(2),* 480-510.

Colman, H. L., & Lunnan, R. (2011). Organizational identification and serendipitous value creation in post-acquisition integration. *Journal of Management, 37(3),* 839-860.

Colquitt, J. A., Long, D. M., Rodell, J. B., & Halvorsen-Ganepola, M. D. (2015). Adding the "in" to justice: A qualitative and quantitative investigation of the differential effects of justice rule adherence and violation. *Journal of Applied Psychology, 100(2),* 278 - 297.

Connelly, C. E., Zweig, D., Webster, J., & Trougakos, J. P. (2012). *Knowledge hiding in organizations. Journal of Organizational Behavior, 33(1),* 64-88.

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

Crayne, M. P., & Hunter, S. T. (2018). Historiometry in organizational science: Renewed attention for an established research method. *Organizational Research Methods, 21(1),* 6-29.

Creed, W. D., DeJordy, R., & Lok, J. (2010). Being the change: Resolving institutional contradiction through identity work. *Academy of Management Journal, 53(6),* 1336-1364.

Crilly, D., Hansen, M., & Zollo, M. (2016). The grammar of decoupling: A cognitive-linguistic perspective on firms' sustainability claims and stakeholders' interpretation. *Academy of Management Journal, 59(2),* 705-729.

Dacin, M. T., Munir, K., & Tracey, P. (2010). Formal dining at Cambridge colleges: Linking ritual performance and institutional maintenance. *Academy of Management Journal, 53(6),* 1393-1418.

Demigha, S. (2016). Mining Knowledge of the Patient Record: The Bayesian classification to predict and detect anomalies in breast cancer. *Electronic Journal of Knowledge Management, 14(3),* 128 - 139.

Doucet, L., & Jehn, K. A. (1997). Analyzing harsh words in a sensitive setting: American expatriates in communist China. *Journal of Organizational Behavior, 18(S1),* 559-582.

Duriau, V. J., Reger, R. K., & Pfarrer, M. D. (2007). A content analysis of the content analysis literature in organization studies: Research themes, data sources, and methodological refinements. *Organizational Research Methods, 10(1),* 5-34.

Elg, U., Ghauri, P. N., Child, J., & Collinson, S. (2017). MNE microfoundations and routines for building a legitimate and sustainable position in emerging markets. *Journal of Organizational Behavior, 38(9),* 1320-1337.

Engelen, A., Neumann, C., & Schmidt, S. (2016). Should entrepreneurially oriented firms have narcissistic CEOs? *Journal of Management, 42(3),* 698-721.

Eury, J. L., Kreiner, G. E., Treviño, L. K., & Gioia, D. A. (2018). The past is not dead: Legacy identification and alumni ambivalence in the wake of the Sandusky scandal at Penn State. *Academy of Management Journal, 61(3),* 826-856.

Felbermayr, A., & Nanopoulos, A. (2016). The role of emotions for the perceived usefulness in online customer reviews. *Journal of Interactive Marketing, 36,* 60-76.

Felps, W., Mitchell, T. R., Hekman, D. R., Lee, T. W., Holtom, B. C., & Harman, W. S. (2009). Turnover contagion: How coworkers' job embeddedness and job search behaviors influence quitting. *Academy of Management Journal, 52(3),* 545-561.

Fenton-O'Creevy, M., Soane, E., Nicholson, N., & Willman, P. (2011). Thinking, feeling and deciding: The influence of emotions on the decision making and performance of traders. *Journal of Organizational Behavior, 32(8),* 1044-1061.

Fisher, K., & Hutchings, K. (2013). Making sense of cultural distance for military expatriates operating in an extreme context. *Journal of Organizational Behavior, 34(6),* 791-812.

Follmer, E. H., Talbot, D. L., Kristof-Brown, A. L., Astrove, S. L., & Billsberry, J. (2018). Resolution, relief, and resignation: A qualitative study of responses to misfit at work. *Academy of Management Journal, 61(2),* 440-465.

Friedman, R., Anderson, C., Brett, J., Olekalns, M., Goates, N., & Lisco, C. C. (2004). The positive and negative effects of anger on dispute resolution: evidence from electronically mediated disputes. *Journal of Applied Psychology, 89(2),* 369 - 376.

Gamache, D. L., & McNamara, G. (in press.) Responding to bad press: How CEO temporal focus influences the sensitivity to negative media coverage of acquisitions. *Academy of Management Journal.*

Gamache, D. L., McNamara, G., Mannor, M. J., & Johnson, R. E. (2015). Motivated to acquire? The impact of CEO regulatory focus on firm acquisitions. *Academy of Management Journal, 58*(4), 1261-1282.

Gan, Q., Ferns, B. H., Yu, Y., & Jin, L. (2017). A text mining and multidimensional sentiment analysis of online restaurant reviews. *Journal of Quality Assurance in Hospitality & Tourism, 1*8(4), 465-492.

Gangloff, K. A., Connelly, B. L., & Shook, C. L. (2016). Of scapegoats and signals: Investor reactions to CEO succession in the aftermath of wrongdoing. *Journal of Management, 42*(6), 1614-1634.

Gelfand, M. J., Severance, L., Lee, T., Bruss, C. B., Lun, J., Abdel-Latif, A. H., ... & Moustafa Ahmed, S. (2015). Culture and getting to yes: The linguistic signature of creative agreements in the United States and Egypt. *Journal of Organizational Behavior, 36*(7), 967-989.

Gephart, R. (1997). Hazardous measures: An interpretive textual analysis of quantitative sensemaking during crises. *Journal of Organizational Behavior, 18*(S1), 583-622.

Gibson, C. B., & Gibbs, J. L. (2006). Unpacking the concept of virtuality: The effects of geographic dispersion, electronic dependence, dynamic structure, and national diversity on team innovation. *Administrative Science Quarterly, 51*(3), 451-495.

Gibson, C. B., & Zellmer-Bruhn, M. E. (2001). Metaphors and meaning: An intercultural analysis of the concept of teamwork. *Administrative Science Quarterly, 46*(2), 274-303.

Gioia, D. A., Price, K. N., Hamilton, A. L., & Thomas, J. B. (2010). Forging an identity: An insider-outsider study of processes involved in the formation of organizational identity. *Administrative Science Quarterly, 55*(1), 1-46.

Godnov, U., & Redek, T. (2018). Good food, clean rooms and friendly staff: Implications of user-generated content for Slovenian skiing, sea and spa hotels' management. *Management: Journal of Contemporary Management Issues, 23(*1), 29-57.

Gomulya, D., & Boeker, W. (2014). How firms respond to financial restatement: CEO successors and external reactions. *Academy of Management Journal, 57(*6), 1759-1785.

Gomulya, D., Wong, E. M., Ormiston, M. E., & Boeker, W. (2017). The role of facial appearance on CEO selection after firm misconduct. *Journal of Applied Psychology, 102*(4), 617-635.

Gover, L., & Duxbury, L. (2018). Making sense of organizational change: Is hindsight really 20/20?. *Journal of Organizational Behavior, 39*(1), 39-51.

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.

Grzywacz, J. G., Arcury, T. A., Marín, A., Carrillo, L., Burke, B., Coates, M. L., & Quandt, S. A. (2007). Work-family conflict: Experiences and health implications among immigrant Latinos. *Journal of Applied Psychology, 92*(4), 1119 - 1130.

Guo, W., Yu, T., & Gimeno, J. (2017). Language and competition: Communication vagueness, interpretation difficulties, and market entry. *Academy of Management Journal, 60*(6), 2073-2098.

Gutierrez, B., Howard-Grenville, J., & Scully, M. A. (2010). The faithful rise up: Split identification and an unlikely change effort. *Academy of Management Journal, 53*(4), 673-699.

Hagtvedt, L. P., Dossinger, K., Harrison, S. H., & Huang, L. (2019). Curiosity made the cat more creative: Specific curiosity as a driver of creativity. *Organizational Behavior and Human Decision Processes, 150,* 1-13.

Hájek, P. (2018). Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns. *Neural Computing and Applications, 29*(7), 343-358.

Hannigan, T., Haans, R. F. J., Vakili, K., Tchalian, H., Glaser, V., Wang, M., ... & Jennings, P. D. (in press). Topic modeling in management research: Rendering new theory from textual data. *Academy of Management Annals*.

Hao, J., & Dai, H. (2016). Social media content and sentiment analysis on consumer security breaches. *Journal of Financial Crime, 23*(4), 855-869.

Harrison, S. H., & Dossinger, K. (2017). Pliable guidance: A multilevel model of curiosity, feedback seeking, and feedback giving in creative work. *Academy of Management Journal, 60*(6), 2051-2072.

He, W., Zhang, W., Tian, X., Tao, R., & Akula, V. (2019). Identifying customer knowledge on social media through data analytics. *Journal of Enterprise Information Management, 32*(1), 152-169.

Heyden, M. L., Sidhu, J. S., & Volberda, H. W. (2018). The conjoint influence of top and middle management characteristics on management innovation. *Journal of Management, 44*(4), 1505-1529.

Hoff, T. J., Pohl, H., & Bartfield, J. (2006). Teaching but not learning: how medical residency programs handle errors. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior, 27*(7), 869-896.

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

Howard-Grenville, J., Metzger, M. L., & Meyer, A. D. (2013). Rekindling the flame: Processes of identity resurrection. *Academy of Management Journal,* 56(1), 113-136.

Hu, Y. H., Chen, Y. L., & Chou, H. L. (2017). Opinion mining from online hotel reviews–A text summarization approach. *Information Processing & Management, 53*(2), 436-449.

Huang, L. (2018). The role of investor gut feel in managing complexity and extreme risk. *Academy of Management Journal, 61*(5), 1821-1847.

Huang, M., ElTayeby, O., Zolnoori, M., & Yao, L. (2018). Public opinions ioward diseases: Infodemiological study on news media data. *Journal of Medical Internet Research, 20*(5), e10047.

Hubbard, T. D., Pollock, T. G., Pfarrer, M. D., & Rindova, V. P. (2018). Safe bets or hot hands? How status and celebrity influence strategic alliance formations by newly public firms. *Academy of Management Journal, 61*(5), 1976-1999.

Janasik, N., Honkela, T., & Bruun, H. (2009). Text mining in qualitative research: Application of an unsupervised learning method. *Organizational Research Methods, 12*(3), 436-460.

Jancenelle, V. E. (2018). Organizational psychological capital during earnings conference calls: Mitigating shareholders' sell-off in the face of earnings surprises? *Journal of Leadership & Organizational Studies, 25*(4), 469-480.

Jancenelle, V. E., & Javalgi, R. R. G. (2018). The effect of moral foundations in prosocial crowdfunding. *International Small Business Journal, 36*(8), 932-951.

Jancenelle, V. E., Javalgi, R. R. G., & Cavusgil, E. (2018). The role of economic and normative signals in international prosocial crowdfunding: An illustration using market orientation and psychological capital. *International Business Review, 27*(1), 208-217.

Jancenelle, V. E., Storrud-Barnes, S. F., & Iaquinto, A. (2019). Making investors feel good during earnings conference calls: The effect of warm-glow rhetoric. *Journal of General Management, 44*(2), 63-72.

Jancenelle, V. E., Storrud-Barnes, S., & Javalgi, R. R. G. (2017). Corporate entrepreneurship and market performance: A content analysis of earnings conference calls. *Management Research Review, 40*(3), 352-367.

Jarvis, L. C., Goodrick, E., & Hudson, B. A. (in press.). Where the heart functions best: Reactive-affective conflict and the disruptive work of animal rights organizations. *Academy of Management Journal*.

Jehn, K. A., & Bezrukova, K. (2004). A field study of group diversity, workgroup context, and performance. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior, 25*(6), 703-729.

Jia, S. (2018). Behind the ratings: Text mining of restaurant customers' online reviews. *International Journal of Market Research, 60*(6), 561-572.

Johnman, M., Vanstone, B. J., & Gepp, A. (2018). Predicting FTSE 100 returns and volatility using sentiment analysis. *Accounting & Finance, 58,* 253-274.

Kabanoff, B., & Holt, J. (1996). Changes in the espoused values of Australian organizations 1986-1990. *Journal of Organizational Behavior, 17*(3), 201-219.

Kabanoff, B., Waldersee, R., & Cohen, M. (1995). Espoused values and organizational change themes. *Academy of Management Journal, 38*(4), 1075-1104.

Kakol, M., Nielek, R., & Wierzbicki, A. (2017). Understanding and predicting Web content credibility using the Content Credibility Corpus. *Information Processing & Management, 53*(5), 1043-1061.

Kananovich, V. (2018). Framing the taxation-democratization link: An automated content analysis of cross-national newspaper data. *The International Journal of Press/Politics, 23*(2), 247-267.

Kang, T., Park, D. H., & Han, I. (2018). Beyond the numbers: The effect of 10-K tone on firms' performance predictions using text analytics. *Telematics and Informatics, 35*(2), 370-381.

Kanze, D., Huang, L., Conley, M. A., & Higgins, E. T. (2018). We ask men to win and women not to lose: Closing the gender gap in startup funding. *Academy of Management Journal, 61*(2), 586-614.

Khan, F. H., Qamar, U., & Bashir, S. (2016). Multi-objective model selection (MOMS)-based semi-supervised framework for sentiment analysis. *Cognitive Computation, 8*(4), 614-628.

Khan, F. H., Qamar, U., & Bashir, S. (2016). Senti-CS: Building a lexical resource for sentiment analysis using subjective feature selection and normalized chi-square-based feature weight generation. *Expert Systems, 33*(5), 489-500.

Kim, E. H. J., Jeong, Y. K., Kim, Y., Kang, K. Y., & Song, M. (2016). Topic-based content and sentiment analysis of ebola virus on twitter and in the news. *Journal of Information Science, 42*(6), 763-781.

King, E. B., Shapiro, J. R., Hebl, M. R., Singletary, S. L., & Turner, S. (2006). The stigma of obesity in customer service: A mechanism for remediation and bottom-line consequences of interpersonal discrimination. *Journal of Applied Psychology, 91*(3), 579 - 593.

Kistruck, G. M., Lount Jr, R. B., Smith, B. R., Bergman Jr, B. J., & Moss, T. W. (2016). Cooperation vs. competition: Alternative goal structures for motivating groups in a resource scarce environment. *Academy of Management Journal, 59* (4), 1174-1198.

Kjell, O. N., Kjell, K., Garcia, D., & Sikström, S. (2018). Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods, 24(1),* 92.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihok, G., & Den Hartog, D. N. (2018). Text classification for organizational researchers: A tutorial. *Organizational Research Methods, 21*(3), 766-799.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihok, G., & Den Hartog, D. N. (2018). Text mining in organizational research. *Organizational Research Methods, 21*(3), 733-765.

König, A., Mammen, J., Luger, J., Fehn, A., & Enders, A. (2018). Silver Bullet or Ricochet? CEOs' Use of Metaphorical Communication and Infomediaries' Evaluations. *Academy of Management Journal, 6*1(4), 1196-1230.

Kregel, I., Ogonek, N., & Matthies, B. (2019). Competency profiles for lean professionals–an international perspective. *International Journal of Productivity and Performance Management, 68*(2), 423-446.

Kreiner, G. E., Hollensbe, E., Sheep, M. L., Smith, B. R., & Kataria, N. (2015). Elasticity and the dialectic tensions of organizational identity: How can we hold together while we are pulling apart?. *Academy of Management Journal, 58*(4), 981-1011.

Kreiner, G. E., Hollensbe, E. C., & Sheep, M. L. (2009). Balancing borders and bridges: Negotiating the work-home interface via boundary work tactics. *Academy of Management Journal, 52*(4), 704-730.

Kreiner, G. E., Hollensbe, E. C., & Sheep, M. L. (2006). Where is the "me" among the "we"? Identity work and the search for optimal balance. *Academy of Management Journal, 49*(5), 1031-1057.

Kwantes, P. J., Derbentseva, N., Lam, Q., Vartanian, O., & Marmurek, H. H. (2016). Assessing the Big Five personality traits with latent semantic analysis. *Personality and Individual Differences, 102,* 229-233.

La Bella, A., Fronzetti Colladon, A., Battistoni, E., Castellan, S., & Francucci, M. (2018). Assessing perceived organizational leadership styles through twitter text mining. *Journal of the Association for Information Science and Technology, 69*(1), 21-31.

Lanaj, K., Foulk, T. A., & Erez, A. (2019). Energizing leaders via self-reflection: A within-person field experiment. *Journal of Applied Psychology, 104*(1), 1-18.

Lawrence, T. B. (1999). Institutional strategy. *Journal of Management, 25*(2), 161-187.

Lawrence, T. B. (2017). High-stakes institutional translation: Establishing North America's first government-sanctioned supervised injection site. *Academy of Management Journal, 60*(5), 1771-1800.

Lehdonvirta, V., Kässi, O., Hjorth, I., Barnard, H., & Graham, M. (2019). The global platform economy: A new offshoring institution enabling emerging-economy microproviders. *Journal of Management, 45*(2), 567-599.

Levy, O. (2005). The influence of top management team attention patterns on global strategic posture of firms. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior, 26*(7), 797-819.

Liang, L. H., Lian, H., Brown, D. J., Ferris, D. L., Hanig, S., & Keeping, L. M. (2016). Why are abusive supervisors abusive? A dual-system self-control model. *Academy of Management Journal, 59*(4), 1385-1406.

Lingo, E. L., & O'Mahony, S. (2010). Nexus work: Brokerage on creative projects. *Administrative Science Quarterly, 55(*1), 47-81.

Lips-Wiersma, M., & Hall, D. T. (2007). Organizational career development is not dead: A case study on managing the new career during organizational change. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior, 28*(6), 771-792.

Lissack, M. R. (1998). Concept sampling: A new twist for content analysis. *Organizational Research Methods, 1*(4), 484-504.

Liu, Y. H., & Chen, Y. L. (2018). A two-phase sentiment analysis approach for judgement prediction. *Journal of Information Science, 44*(5), 594-607.

Liu, X., Burns, A. C., & Hou, Y. (2017). An investigation of brand-related user-generated content on Twitter. *Journal of Advertising, 46(2),* 236-247.

Love, E. G., Lim, J., & Bednar, M. K. (2017). The face of the firm: The influence of CEOs on corporate reputation. *Academy of Management Journal, 60*(4), 1462-1481.

Luciano, M. M., Mathieu, J. E., Park, S., & Tannenbaum, S. I. (2018). A fitting approach to construct and measurement alignment: The role of big data in advancing dynamic theories. *Organizational Research Methods, 21*(3), 592-632.

Madera, J. M., Hebl, M. R., & Martin, R. C. (2009). Gender and letters of recommendation for academia: agentic and communal differences. *Journal of Applied Psychology, 94*(6), 1591.

Mair, J., Marti, I., & Ventresca, M. J. (2012). Building inclusive markets in rural Bangladesh: How intermediaries work institutional voids. *Academy of Management Journal, 55*(4), 819-850.

Malhotra, S., Reus, T. H., Zhu, P., & Roelofsen, E. M. (2018). The acquisitive nature of extraverted CEOs. *Administrative Science Quarterly, 63*(2), 370-408.

Mantere, S., Schildt, H. A., & Sillince, J. A. (2012). Reversal of strategic change. *Academy of Management Journal, 55*(1), 172-196.

Martin, S. R. (2016). Stories about values and valuable stories: A field experiment of the power of narratives to shape newcomers' actions. *Academy of Management Journal, 59*(5), 1707-1724.

Martin, A. E., & Phillips, K. W. (2017). What "blindness" to gender differences helps women see and do: Implications for confidence, agency, and action in male-dominated environments. *Organizational Behavior and Human Decision Processes, 142,* 28-44.

Mazis, P., & Tsekrekos, A. (2017). Latent semantic analysis of the FOMC statements. *Review of Accounting and Finance, 16*(2), 179-217.

McAlearney, A. S. (2006). Leadership development in healthcare: a qualitative study. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior, 27*(7), 967-982.

McClelland, P. L., Liang, X., & Barker III, V. L. (2010). CEO commitment to the status quo: Replication and extension using content analysis. *Journal of Management, 36(*5), 1251-1277.

McKenny, A. F., Short, J. C., & Payne, G. T. (2013). Using computer-aided text analysis to elevate constructs: An illustration using psychological capital. *Organizational Research Methods, 16*(1), 152-184.

McKenny, A. F., Aguinis, H., Short, J. C., & Anglin, A. H. (2018). What doesn't get measured does exist: Improving the accuracy of computer-aided text analysis. *Journal of Management, 44*(7), 2909-2933.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 5(4),* 1093-1113.

Menon, A., Choi, J., & Tabakovic, H. (2018, July). What you say your strategy is and why it matters: Natural language processing of unstructured text. In *Academy of Management Proceedings* (Vol. 2018, No. 1, p. 18319). Briarcliff Manor, NY 10510: Academy of Management.

Moore, C., Lee, S. Y., Kim, K., & Cable, D. M. (2017). The advantage of being oneself: The role of applicant self-verification in organizational hiring decisions. *Journal of Applied Psychology, 102*(11), 1493-1513.

Morris, R. (1994). Computerized content analysis in management research: A demonstration of advantages & limitations. *Journal of Management, 20(4),* 903 - 931.

Mossholder, K. W., Settoon, R. P., Harris, S. G., & Armenakis, A. A. (1995). Measuring emotion in open-ended survey responses: An application of textual data analysis. *Journal of Management, 21(2),* 335-355.

Nadkarni, S., & Chen, J. (2014). Bridging yesterday, today, and tomorrow: CEO temporal focus, environmental dynamism, and rate of new product introduction. *Academy of Management Journal, 57*(6), 1810-1833.

Nadkarni, S., & Narayanan, V. K. (2005). Validity of the structural properties of text-based causal maps: An empirical assessment. *Organizational Research Methods, 8*(1), 9-40.

Nadkarni, S., Pan, L., & Chen, T. (2019). Only timeline will tell: Temporal framing of competitive announcements and rivals' responses. *Academy of Management Journal, 62*(1), 117-143.

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.

Nakayama, M., & Wan, Y. (2019). The cultural impact on social commerce: A sentiment analysis on Yelp ethnic restaurant reviews. *Information & Management, 56*(2), 271-279.

Navis, C., & Glynn, M. A. (2010). How new market categories emerge: Temporal dynamics of legitimacy, identity, and entrepreneurship in satellite radio, 1990–2005. *Administrative Science Quarterly, 55*(3), 439-471.

Neeley, T. B., & Dumas, T. L. (2016). Unearned status gain: Evidence from a global language mandate. *Academy of Management Journal, 59*(1), 14-43.

Nimon, K., Shuck, B., & Zigarmi, D. (2016). Construct overlap between employee engagement and job satisfaction: A function of semantic equivalence? *Journal of Happiness Studies, 17*(3), 1149-1171.

Olsen, A. Ø., Sofka, W., & Grimpe, C. (2016). Coordinated exploration for grand challenges: The role of advocacy groups in search consortia. *Academy of Management Journal, 59*(6), 2232-2255.

Onan, A., & Korukoğlu, S. (2017). A feature selection model based on genetic rank aggregation for text sentiment classification. *Journal of Information Science, 43*(1), 25-38.

Owens, B. P., & Hekman, D. R. (2012). Modeling how to grow: An inductive examination of humble leader behaviors, contingencies, and outcomes. *Academy of Management Journal, 55*(4), 787-818.

Ozcan, P., & Gurses, K. (2018). Playing ccat and mouse: Contests over regulatory categorization of dietary supplements in the United States. *Academy of Management Journal, 61*(5), 1789-1820.

Öztürk, N., & Ayvaz, S. (2018). Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. *Telematics and Informatics, 35*(1), 136-147.

Palmer, I., Kabanoff, B., & Dunford, R. (1997). Managerial accounts of downsizing. *Journal of Organizational Behavior, 18*(S1), 623-639.

Pandey, S., & Pandey, S. K. (2017). Applying natural language processing capabilities in computerized textual analysis to measure organizational culture. *Organizational Research Methods, 22(3),* 1094428117745648.

Park, M. (2019). What's Important: An exploratory analysis of student evaluations about physics professors on RateMyProfessors. com. *Journal of College Science Teaching, 48*(4), 36-44.

Park, S. H., & Hong, S. H. (2018). Identification of primary medication concerns regarding thyroid hormone replacement therapy from online patient medication reviews: Text mining of social network data. *Journal of Medical Internet Research, 20*(10), e11085.

Pencle, N., & Mălăescu, I. (2016). What's in the words? Development and validation of a multidimensional dictionary for CSR and application using prospectuses. *Journal of Emerging Technologies in Accounting, 13(2),* 109-127.

Owens, B. P., & Hekman, D. R. (2012). Modeling how to grow: An inductive examination of humble leader behaviors, contingencies, and outcomes. *Academy of Management Journal, 55*(4), 787-818.

Piezunka, H., & Dahlander, L. (2019). Idea rejected, tie formed: organizations' feedback on crowdsourced ideas. *Academy of Management Journal, 62(2),* 503-530.

Pollach, I. (2012). Taming textual data: The contribution of corpus linguistics to computer-aided text analysis. *Organizational Research Methods, 15(2),* 263-287.

Porac, J. F., Wade, J. B., & Pollock, T. G. (1999). Industry categories and the politics of the comparable firm in CEO compensation. *Administrative Science Quarterly, 44(*1), 112-144.

Petriglieri, J. L. (2015). Co-creating relationship repair: Pathways to reconstructing destabilized organizational identification. *Administrative Science Quarterly, 60*(3), 518-557.

Quigley, T. J., Hubbard, T. D., Ward, A., & Graffin, S. D. (in press.). Unintended consequences: Information releases and CEO stock option grants. *Academy of Management Journal.*

Raffaelli, R. (in press.). Technology Reemergence: Creating New Value for old technologies in Swiss mechanical watchmaking, 1970-2008. *Administrative Science Quarterly,* 1 - 43

Ramarajan, L., Bezrukova, K., Jehn, K. A., & Euwema, M. (2011). From the outside in: The negative spillover effects of boundary spanners' relations with members of other organizations. *Journal of Organizational Behavior, 32*(6), 886-905.

Reay, T., Goodrick, E., Waldorff, S. B., & Casebeer, A. (2017). Getting leopards to change their spots: Co-creating a new professional role identity. A*cademy of Management Journal, 60*(3), 1043-1070.

Reinecke, J., & Ansari, S. (2015). When times collide: Temporal brokerage at the intersection of markets and developments. *Academy of Management Journal, 58*(2), 618-648.

Rhee, E. Y., & Fiss, P. C. (2014). Framing controversial actions: Regulatory focus, source credibility, and stock market reaction to poison pill adoption. *Academy of Management Journal, 57*(6), 1734-1758.

Ridge, J. W., & Ingram, A. (2017). Modesty in the top management team: Investor reaction and performance implications. *Journal of Management, 43*(4), 1283-1306.

Rothausen, T. J., Henderson, K. E., Arnold, J. K., & Malshe, A. (2017). Should I stay or should I go? Identity and well-being in sensemaking about retention and turnover. *Journal of Management, 43*(7), 2357-2385.

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (in press). Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology.*

Sandhu, S., & Kulik, C. T. (in press). Shaping and Being Shaped: How Oorganizational structure and Mmanagerial discretion co-evolve in new managerial roles *Administrative Sciene Quarterly*.

Santistevan, D., & Josserand, E. (2019). Meta-teams: Getting global work done in MNEs. *Journal of Management, 45*(2), 510-539.

Savani, K., & King, D. (2015). Perceiving outcomes as determined by external forces: The role of event construal in attenuating the outcome bias. *Organizational Behavior and Human Decision Processes, 130,* 136-146.

Saxena, A., Chaturvedi, K. R., & Rakesh, S. (2018). Analysing customers reactions on social media promotional campaigns: A text-mining approach. *Paradigm, 22*(1), 80-99.

Sbalchiero, S., & Tuzzi, A. (2016). Scientists' spirituality in scientists' words. Assessing and enriching the results of a qualitative analysis of in-depth interviews by means of quantitative approaches. *Quality & Quantity, 50*(3), 1333-1348.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one, 8*(9), e73791.

Scott, D., Oh, J., Chappelka, M., Walker-Holmes, M., & DiSalvo, C. (2018). Food for Thought: analyzing public opinion on the supplemental nutrition assistance program. *Journal of Technology in Human Services, 36(*1), 37-47.

Sgourev, S. V., & Operti, E. (in press). From Montagues To Capulets? Analyzing the Systemic Nature of Rivalry in Career Mobility. *Academy of Management Journal.*

Shantz, A., & Latham, G. P. (2009). An exploratory field experiment of the effect of subconscious and conscious goals on employee performance. *Organizational Behavior and Human Decision Processes, 109*(1), 9-17.

Shi, W., Zhang, Y., & Hoskisson, R. E. (2019). Examination of CEO–CFO social interaction through language style matching: Outcomes for the CFO and the organization. *Academy of Management Journal, 62*(2), 383-414.

Short, J. C., & Palmer, T. B. (2008). The application of DICTION to content analysis research in strategic management. *Organizational Research Methods, 11*(4), 727-752.

Short, J. C., Broberg, J. C., Cogliser, C. C., & Brigham, K. H. (2010). Construct validation using computer-aided text analysis (CATA) an illustration using entrepreneurial orientation. *Organizational Research Methods, 13*(2), 320-347.

Short, J. C., McKenny, A. F., & Reid, S. W. (2018). More than words? Computer-aided text analysis in organizational behavior and psychology research. *Annual Review of Organizational Psychology and Organizational Behavior, 5,* 415-435.

Shortt, H. L., & Warren, S. K. (2019). Grounded visual pattern analysis: Photographs in organizational field studies. *Organizational Research Methods, 22*(2), 539-563.

Slutskaya, N., Game, A. M., & Simpson, R. C. (2018). Better together: Examining the role of collaborative ethnographic documentary in organizational research. *Organizational Research Methods, 21*(2), 341-365.

Smets, M., Jarzabkowski, P., Burke, G. T., & Spee, P. (2015). Reinsurance trading in Lloyd's of London: Balancing conflicting-yet-complementary logics in practice. *Academy of management journal, 58*(3), 932-970.

Smither, J. W., & Walker, A. G. (2004). Are the characteristics of narrative comments related to improvement in multirater feedback ratings over time?. *Journal of Applied Psychology, 89*(3), 575.

Sonenshein, S. (2014). How organizations foster the creative use of resources. *Academy of Management Journal, 57*(3), 814-848.

Sorour, S. E., Goda, K., & Mine, T. (2017). Comment data mining to estimate student performance considering consecutive lessons. *Journal of Educational Technology & Society, 20*(1), 73.

Speer, A. B. (2018). Quantifying with words: An investigation of the validity of narrative-derived performance scores. *Personnel Pscyhology, 71(3),* 299-333.

Spell, C. S., & Blum, T. C. (2005). Adoption of workplace substance abuse prevention programs: Strategic choice and institutional perspectives. *Academy of Management Journal, 48*(6), 1125-1142.

Srinivas, S., & Rajendran, S. (2019). Topic-based knowledge mining of online student reviews for strategic planning in universities. *Computers & Industrial Engineering, 128,* 974-984.

Stanko, T. L., & Beckman, C. M. (2015). Watching you watching me: Boundary control and capturing attention in the context of ubiquitous technology use. *Academy of Management Journal, 58*(3), 712-738.

Stuckman, J., Walden, J., & Scandariato, R. (2017). The effect of dimensionality reduction on software vulnerability prediction models. *IEEE Transactions on Reliability, 66(*1), 17-37.

Tilcsik, A. (2010). From ritual to reality: Demography, ideology, and decoupling in a post-communist government agency. *Academy of Management Journal, 53*(6), 1474-1498.

Titus Jr, V., House, J. M., & Covin, J. G. (2017). The influence of exploration on external corporate venturing activity. *Journal of Management, 43*(5), 1609-1630.

Titus Jr, V. K., Parker, O., & Bass, A. E. (2018). Ripping off the band-aid: Scrutiny bundling in the wake of social disapproval. *Academy of Management Journal, 61*(2), 637-660.

Titus Jr, V. K., Parker, O., & Bass, A. E. (2018). Ripping off the band-aid: Scrutiny bundling in the wake of social disapproval. *Academy of Management Journal, 61*(2), 637-660.

Toegel, G., Kilduff, M., & Anand, N. (2013). Emotion helping by managers: An emergent understanding of discrepant role expectations and outcomes. *Academy of Management Journal, 56*(2), 334-357.

Treviño, L. K., den Nieuwenboer, N. A., Kreiner, G. E., & Bishop, D. G. (2014). Legitimating the legitimate: A grounded theory study of legitimacy work among Ethics and Compliance Officers. *Organizational Behavior and Human Decision Processes, 123*(2), 186-205.

Tsukioka, Y., Yanagi, J., & Takada, T. (2018). Investor sentiment extracted from internet stock message boards and IPO puzzles. *International Review of Economics & Finance, 56,* 205-217.

Tudoran, A. A. (2019). Why do internet consumers block ads? New evidence from consumer opinion mining and sentiment analysis. *Internet Research, 29*(1), 144-166.

Usai, A., Pironti, M., Mital, M., & Aouina Mejri, C. (2018). Knowledge discovery out of text data: a systematic review via text mining. *Journal of Knowledge Management, 22*(7), 1471-1488.

Vagnani, G. (2015). Exploration and long-run organizational performance: the moderating role of technological interdependence. *Journal of Management, 4*1(6), 1651-1676.

Valdez, D., Pickett, A. C., & Goodson, P. (2018). Topic modeling: Latent semantic analysis for the social sciences. *Social Science Quarterly, 99*(5), 1665-1679.

Valentine, M. A. (2018). Renegotiating spheres of obligation: The role of hierarchy in organizational learning. *Administrative Science Quarterly, 63*(3), 570-606.

Van Dijke, M., Wildschut, T., Leunissen, J. M., & Sedikides, C. (2015). Nostalgia buffers the negative impact of low procedural justice on cooperation. *Organizational Behavior and Human Decision Processes, 127,* 15-29.

Van Wijk, J., Stam, W., Elfring, T., Zietsma, C., & Den Hond, F. (2013). Activists and incumbents structuring change: The interplay of agency, culture, and networks in field evolution. *Academy of Management Journal, 56*(2), 358-386.

Vergne, J. P., & Depeyre, C. (2016). How do firms adapt? A fuzzy-set analysis of the role of cognition and capabilities in US defense firms' responses to 9/11. *Academy of Management Journal, 59*(5), 1653-1680.

Vracheva, V., Judge, W. Q., & Madden, T. (2016). Enterprise strategy concept, measurement, and validation: Integrating stakeholder engagement into the firm's strategic architecture. *European Management Journal, 34*(4), 374-385.

Wade, J. B., Porac, J. F., & Pollock, T. G. (1997). Worth, words, and the justification of executive pay. *Journal of Organizational Behavior, 18*(S1), 641-664.

Walker, D. D., van Jaarsveld, D. D., & Skarlicki, D. P. (2017). Sticks and stones can break my bones but words can also hurt me: The relationship between customer verbal aggression and employee incivility. *Journal of Applied Psychology, 102*(2), 163 - 179.

Waller, M. J., & Kaplan, S. A. (2018). Systematic behavioral observation for emergent team phenomena: Key considerations for quantitative video-based approaches. *Organizational Research Methods, 21*(2), 500-515.

Wang, T., Wezel, F. C., & Forgues, B. (2016). Protecting market identity: When and how do organizations respond to consumers' devaluations?. *Academy of Management Journal, 59*(1), 135-162.

Wen, Q., Qiang, M., Xia, B., & An, N. (in press.). Discovering regulatory concerns on bridge management: An author-topic model based approach. *Transport Policy.*

Distribution A. Approved for public release; distribution unlimited.
88ABW-2019-5796; Cleared 05 December 2019

Whiteman, G., & Cooper, W. H. (2011). Ecological sensemaking. *Academy of Management Journal, 54*(5), 889-911.

Wiedner, R., & Mantere, S. (in press.). Cutting the cord: mutual respect, organizational autonomy, and independence in organizational separation processes. *Administrative Science Quarterly.*

Wiedner, R., Barrett, M., & Oborn, E. (2017). The emergence of change in unexpected places: Resourcing across organizational practices in strategic change. *Academy of Management Journal, 60(3),* 823-854.

Wilhelmy, A., Kleinmann, M., König, C. J., Melchers, K. G., & Truxillo, D. M. (2016). How and why do interviewers try to make impressions on applicants? A qualitative study. *Journal of Applied Psychology, 101*(3), 313 - 332.

Williams, T. A., & Shepherd, D. A. (2017). Mixed method social network analysis: Combining inductive concept development, content analysis, and secondary data for quantitative analysis. *Organizational Research Methods, 20*(2), 268-298.

Wilson, K. S., DeRue, D. S., Matta, F. K., Howe, M., & Conlon, D. E. (2016). Personality similarity in negotiations: Testing the dyadic effects of similarity in interpersonal traits and the use of emotional displays on negotiation outcomes. *Journal of Applied Psychology, 101*(10), 1405 - 1421.

Wolfe, R. A., Gephart, R. P., & Johnson, T. E. (1993). Computer-facilitated qualitative data analysis: Potential contributions to management research. *Journal of Management, 19*(3), 637-660.

Wright, C., & Nyberg, D. (2017). An inconvenient truth: How organizations translate climate change into business as usual. *Academy of Management Journal, 60*(5), 1633-1661.

Wright, A. L., Zammuto, R. F., & Liesch, P. W. (2017). Maintaining the values of a profession: Institutional work and moral emotions in the emergency department. *Academy of Management Journal, 60*(1), 200-237.

Xie, K., Di Tosto, G., Lu, L., & Cho, Y. S. (2018). Detecting leadership in peer-moderated online collaborative learning through text mining and social network analysis. *The Internet and Higher Education, 38,* 9-17.

Xu, X. (2018). Does traveler satisfaction differ in various travel group compositions? Evidence from online reviews. International *Journal of Contemporary Hospitality Management, 30*(3), 1663-1685.

Xu, Z., & Guo, H. (2018). Using Text Mining to compare online pro-and anti-vaccine headlines: Word usage, sentiments, and online popularity. *Communication Studies, 69*(1), 103-122.

Yahav, I., Shehory, O., & Schwartz, D. (2019). Comments mining with TF-IDF: The inherent bias and its removal. *IEEE Transactions on Knowledge and Data Engineering, i1*(3), 437-450.

Yang, R., Yu, Y., Liu, M., & Wu, K. (2018). Corporate risk disclosure and audit fee: a text mining approach. *European Accounting Review, 27(3),* 583-594.

Yang, H. C., Lee, C. H., & Wu, C. Y. (2018). Sentiment discovery of social messages using self-organizing maps. *Cognitive Computation, 10*(6), 1152-1166.

Yong, S. H. I., Tang, Y. R., Cui, L. X., & Wen, L. O. N. G. (2018). A text mining based study of investor sentiment and its influence on stock returns. *Economic Computation & Economic Cybernetics Studies & Research, 52*(1).

York, J. G., Hargrave, T. J., & Pacheco, D. F. (2016). Converging winds: Logic hybridization in the Colorado wind energy field. *Academy of Management Journal, 59(2),* 579-610.

Zavyalova, A., Pfarrer, M. D., Reger, R. K., & Shapiro, D. L. (2012). Managing the message: The effects of firm actions and industry spillovers on media coverage following wrongdoing. *Academy of Management Journal, 55(5),* 1079-1101.

Zellmer-Bruhn, M., & Gibson, C. (2006). Multinational organization context: Implications for team learning and performance. *Academy of Management Journal, 49(3),* 501-518.

Zundel, M., MacIntosh, R., & Mackay, D. (2018). The utility of video diaries for organizational research. *Organizational Research Mthods, 21(2),* 386-411.

Zuzul, T. (in press). "Matter Battles:" Boundary Objects and the Failure of Collaboration in Two Smart Cities. *Academy of Management Journal.*

APPENDIX B: Links to Additional Resources

Videos

Machine learning (general)

https://www.youtube.com/watch?v=z-EtmaFJieY

https://www.youtube.com/watch?v=cfj6yaYE86U

https://www.youtube.com/watch?v=ukzFI9rgwfU

Unsupervised machine learning

https://www.youtube.com/watch?v=IUn8k5zSI6g

Supervised machine learning

https://www.youtube.com/watch?v=jmLid2x9eKg&t=305s

Websites

Towards Data Science is a great website full of information about machine learning: https://towardsdatascience.com/machine-learning/home

Machine Learning Mastery is a good go-to for quick and dirty explanations on this complex material: https://machinelearningmastery.com/. Also access it here https://machinelearningmastery.com/start-here/ and look at the "Topics" drop-down menu for a selection.

Podcasts

Data Skeptic: https://dataskeptic.com/

Learning Machines 101: https://www.learningmachines101.com/

Talking Machines: https://www.thetalkingmachines.com/

They also have a great website with articles and white papers on machine learning and AI.

Supplemental Excell Spreadsheet with Detailed Reference Material

Campion, M. A., & Campion, E. D. (2019). *Literature review of computer-assisted text analysis reesearch, software, analytical techniques, and best practices: Supplemental spreadsheet with detailed reference material,* AFRL-RH-WP-TR-2019-xxx. Wright-Patterson AFB, Warfighter Interfaces Division, Collaborative Interfaces and Teaming Branch. (Attached to this report)

Click the paper clip in the panel to the left to open the spreadsheet in Excel.. If the paper clip is not visible, click on the left-facing arrow to the left to view the paper clip.

.

LIST OF ACROYNYMS

AERA	American Educational Research Association
CATA	Computer-Assisted Text Analysis
CEO	Chief Executive Officer
DAL	Dictionary of Affect in Language
DLA	Differential Language Analysis
EEOC	Equal Employment Opportunity Commission
ERIC	Education Resources Information Center
FOIA	Freedom of Information Act
HR	Human Resource
IM	Impression Management
IT	InformationTechnology
KSAO	Knowledge, Skills, Abilities, and other Characteristics
LDA	Latent Dirichlet Allocation
LIWC	Linguistic Inquiry and Word Count
LSA	Latent Semantic Analysis
ML	Machine Learning
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
pLSA	Probabilistic Latent Semantic Analysis
RFP	Request for Proposal
RJP	Realistic Job Preview
RPA	Remotely Piloted Aircraft
SAS	Statistical Analysis System
SEC	Securities and Exchange Commission
SIOP	Sheltered Instruction Observation Protocol
SME	Subject Matter Expert
SO	Sensor Operator
SPSS	Statistical Package for Social Sciences
VADER	Valence Aware Dictionary and sEntiment Reasoner