2022

Development and Validation of a Counterproductive Work Behavior Situational Judgment Test With an Open-ended Response Format: A Computerized Scoring Approach

Saba Tavoosi
University of Central Florida

DEVELOPMENT AND VALIDATION OF A COUNTERPRODUCTIVE WORK BEHAVIOR SITUATIONAL JUDGMENT TEST WITH AN OPEN-ENDED RESPONSE FORMAT: A COMPUTERIZED SCORING APPROACH

By

SABA TAVOOSI
B.A. SAN DIEGO STATE UNIVERSITY, 2017

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science in Industrial and Organizational Psychology
in the Department of Psychology
in the College of Sciences
at the University of Central Florida
Orlando, Florida

Spring Term
2022

Major Professor: Shiyang Su

ABSTRACT

Due to the many detrimental effects of counterproductive work behavior (CWB), it is important to measure the construct accurately. Despite this, there are some limitations inherent to current CWB measures that are somewhat problematic, including that they contain items that do not apply to all jobs while missing items that are important for other jobs (Bowling & Gruys, 2010). The current study tackles these issues by drawing on the benefits associated with open-ended response situational judgment tests (SJTs), such as them having the potential for more insight from respondents (Finch et al., 2018), to develop an open-ended response CWB SJT. To minimize the drawbacks currently associated with the manual analysis of open-ended response SJTs (e.g., being time-consuming and costly)—which is also a reason why they are rarely used—the study leverages natural language processing and machine learning to measure CWB. Using a two-dimensional conceptualization of CWB, including CWB against the organization (CWB-O) and individuals (CWB-I), the CWB SJT dimensions had a moderate to strong correlation with the popular CWB scale the Workplace Deviance scale (Bennett & Robinson, 2000). Findings further indicate the CWB SJT to be related to variables typically associated with CWB tendencies, such as neuroticism and trait self-control. By using topic modeling, it was also found that topic prevalence was largely consistent through time both for the full CWB SJT and for individual items, implying the test-retest reliability. The CWB SJT along with R code for analyzing the open-ended responses is provided. Implication of the CWB SJT for research and practice are discussed.

ACKNOWLEDGMENTS

I want to thank Dr. Shiyang Su for her consistent guidance throughout writing this thesis and funding the data collection, Drs. Steve Jex and Haylee Min for their helpful recommendations, Ann Schlotzhauer and Matthew Ng for their help in revising and developing the situational judgment items, and all aforementioned people as well as Moses Rivera, Emily Hill, and Kenzie Dye for their support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

Counterproductive work behavior (CWB) refers to volitional behaviors employees engage in that harm organizations, stakeholders, or both (Spector & Fox, 2005). Examples of these acts include abuse, theft, withdrawal, production deviance, and sabotage (Spector et al., 2006). These acts have been indicated to be responses to various factors such as workplace stressors (e.g., role ambiguity, role conflict), job dissatisfaction (Chen & Spector, 1992), perceived organizational injustice (Berry et al., 2007), as well as consequences of dispositional factors (e.g., the Big Five, trait anger, trait anxiety, locus of control, narcissism; Spector & Fox, 2005). The negative impact of these behaviors affects both organizations and their employees. For example, theft has been estimated to cost organizations billions of dollars in losses every year worldwide (Bennett et al., 2019), something that has also been associated with business failure (Camara & Schneider, 1994; Greenberg, 1990). CWBs have also been associated with reduced productivity and reduced organizational reputation (Dunlop & Lee, 2004). In a similar vein, employees who are targets of CWB have been indicated to have decreased job satisfaction, increased job stress, increased turnover intentions (Budd et al., 1996), and reduced mental and physical well-being (Bowling & Beehr, 2006).

Due to the many detrimental effects of CWB, it is important to measure the construct accurately. Although multiple measures of CWB exist (e.g., Spector et al., 2006; Bennett & Robinson, 2000) and are being used in the current literature, they suffer from some limitations. More specifically, these scales—which are assumed to be generic and appropriate across jobs (Bowling & Gruys, 2010)—include items that are not relevant to some occupations (e.g., “Falsified a receipt to get reimburse for more money than you spent on business expenses”; Bennett & Robinson, 2000; “Put in to be paid for more hours than you worked”; Spector et al.,

2006) while omitting items that are important to other jobs (e.g., falsifying data as a researcher, or taking bribes as a politician; Bowling & Gruys, 2010). Together these limitations can contribute to underestimations of CWB (Bowling & Gruys, 2010). This study attempts to address this problem by developing a constructed response measurement in the form of a situational judgment test (SJT) where respondents can freely disclose their behaviors without being constrained to a numerical scale or pre-made list of response options.

Situational judgment tests (SJTs) are a type of low-fidelity job simulation (Motowidlo et al., 1990) where respondents are given hypothetical work-related scenarios and have to make judgments on the most appropriate responses (McDaniel et al., 2007; McDaniel et al., 2001). Research has shown that SJTs are effective (Whetzel et al., 2008; McDaniel & Nguyen, 2001), valid at predicting job performance (Whetzel & McDaniel, 2009; McDaniel et al., 2001; McDaniel et al., 2007), offer incremental validity over cognitive ability, the Big 5, and the composite of the two measures (McDaniel et al., 2007), are associated low adverse impact (Whetzel et al., 2008; Chan & Schmitt, 1997; Motowidlo & Tippins, 1993), are less prone to faking and coaching than personality measures (Hooper et al., 2006; Robie et al., 2010), and have good applicant perceptions (Kanning et al., 2006). These benefits, along with them being relatively cheap and quick at administering and scoring (Lievens et al., 2019), make SJTs an attractive assessment tool and an alternative for traditional-self report measures.

SJTs can be administered in different formats (e.g., written, oral, video) and include different response options (e.g., multiple-choice, open-response, role-playing). Research has indicated that constructed response format SJTs have multiple benefits over close-ended formats, including lower adverse impact (Lievens et al., 2019; Edward & Arthur, 2007), higher criterion-related validity (Funke & Schuler, 1998), higher ecological validity (Kjell et al., 2019), and potential for more insight from respondents (Finch et al., 2018). Nevertheless, they are rarely used due to their analytical difficulty. In response to this, some researchers have suggested future research to investigate computerized text analysis to analyze open-ended response SJTs (Lievens et al., 2019).

As of recent, computerized text analysis that draws on techniques developed in machine learning and natural language processing (NLP) has been gaining increasing attention in organizational research (e.g., Kobayashi et al., 2018a, 2018b; Schmiedel et al., 2019; Pandey & Pandey, 2019; Hickman et al., 2020) as an alternative and more efficient way of analyzing text data compared to manual coding. Research has indicated that computerized text analysis reduces rater bias and error, is faster, more cost-effective (Roberts et al., 2014; Downer et al., 2019), and less labor-intensive than manual coding (Kobayashi et al., 2018a). These benefits are even further amplified with increased samples and text data (Campion et al., 2016). Further, Campion et al. (2016) showed that the use of computerized text analysis to score accomplishment records had comparable levels of validity and reliability as human raters scoring the same records. This study will capitalize on these advantages by utilizing computerized scoring to analyze the constructed responses of the CWB SJT. More specifically, *n*-grams—a way to segment text responses into predefined units—will be used in combination with machine learning to predict scores on CWB toward an organization (CWB-O) and CWB towards individuals (CWB-I).

This study has several contributions. First, while previous construct-specific SJTs have focused on measuring neutral or positive constructs and outcomes (e.g., team roles, personal initiative, goal orientation, integrity, and emotional intelligence; Mumford et al. 2008; Bledlow & Frese, 2009; Westring et al., 2009; Becker, 2005; Sharma et al., 2013), this study makes a first attempt at using SJTs to measure negative behavior (i.e., CWB). Second, it attempts to address a longstanding issue with current CWB measures that make their utility in certain jobs limited by developing a constructed response CWB SJT where respondents can freely write out their answers without being constrained to response options predetermined by the researcher. Third, it takes an interdisciplinary approach and integrates a test from the industrial and organizational psychology literature (i.e., SJT) with a computerized scoring method from the data science literature, ultimately drawing on the benefits of open-response SJTs while minimizing the costs associated with their manual coding. *R* (R Core Team, 2021) code is also be provided for analyzing the CWB SJT so that other researchers can easily implement the measure and score the responses.

CHAPTER TWO: LITERATURE REVIEW

Counterproductive Work Behavior

Dimensionality

Early research on CWB focused on examining individual deviant behaviors separately instead of as part of a general higher-order construct. These studies examined narrow constructs such as lateness (Blau, 1995), theft (Greenberg, 1990), sabotage (Mangione & Quinn, 1975), workplace violence (Budd et al., 1996), and mobbing (Zapf et al., 1996). Although these behaviors are sometimes still studied in isolation, contemporary literature has shifted towards treating these different deviant behaviors as part of a broader CWB construct. More specifically, some of the literature tends to conceptualize CWB as a hierarchically organized construct with a general factor of CWB at the top, grouping factors (e.g., interpersonal- and organizational deviance; Bennett & Robinson, 2000) in the middle, and specific CWB behaviors at the bottom (Berry, Carpenter, & Barratt, 2012; Berry, Ones, & Sackett, 2007; Sackett & DeVore, 2001). This integration of counterproductive work behaviors into a general construct has also been supported by meta-analytic findings that have indicated it to be related to various variables such as job satisfaction, organizational commitment, organizational justice, the Big Five personality dimensions, and affect (Berry et al., 2007; Dalal et al., 2005). While current literature tends to agree on the conceptualization of CWB as a broad and hierarchical construct, there is a debate regarding the number and type of dimensions of CWB.

Two-dimensional Model. The perhaps most popular model, proposed by Robinson and Bennett (1995), divides CWB into two subcategories: organizational deviance (OD)—which refers to deviance targeted at the organization (e.g., stealing workplace supplies, taking a lot of breaks, being absent)—and interpersonal deviance (ID)—which refers to deviance targeted at individuals within the organization (e.g., stealing from coworkers, aggression toward coworkers, bullying). Bennett and Robinson (2000) argue that the distinction between targets of deviant behavior is important as the same behavior (e.g., theft) targeted towards the organization versus individuals within the organization can result from different antecedents. This has been supported by meta-analytic findings that have indicated OD and ID to be differentially related to the Big Five personality factors of conscientiousness and agreeableness and with some OCB facets (Berry et al., 2007). In further support of the OD and ID distinction, Robinson and Bennett's (1995) multidimensional scaling results indicated that a two-dimensional solution of CWB provided a better fit for their data compared to a one-dimensional solution, and that three-, four-, and five-dimensional solutions did not significantly improve the fit. Based on this, they recommended the two-dimensional model as the most parsimonious and best-fitting model for deviant behavior. In line with this, Bennett and Robinson (2000) found factor-analytic support for the distinction between the OD and ID dimensions. Overall, the categorization of CWBs into behaviors directed toward the organization versus individuals has received support both through the reliability and validity of the workplace deviance measure (Bennett & Robinson, 2000) as well as through their differential situational and dispositional antecedents (Berry et al., 2007), which lend support to the utility of distinguishing between organizational and interpersonal CWBs within the CWB construct.

Five-dimensional Model. In contrast to the two-dimensional conceptualization of CWB, Spector et al. (2006) proposed and developed a counterproductive work behavior checklist (CWB-C) measure of a five-dimensional taxonomy consisting of the dimensions of abuse, theft, withdrawal, production deviance, and sabotage. They also developed an extended version that conceptualized CWB into behaviors directed towards the organization (CWB-O) versus people (CWB-P), similar to Robinson and Bennett's (1995) conceptualization. Their findings suggested that the five proposed dimensions had largely differential antecedents, while there were smaller differences between the CWB-O and CWB-P dimensions. This suggests that the more discriminant five-dimension conceptualization might be better able to capture and explain the antecedents of CWB. This is, in part, supported by Bolton et al.'s (2010) findings that the five-dimensional CWB-C has, to some extent, differential Big Five personality antecedents that the two-dimensional model might mask. For example, they found that while low conscientiousness was related to CWB-O, it was only significantly associated with sabotage and withdrawal (and not production deviance and theft directed at the organization, which are also facets of CWB-O), indicating that the five-dimensional model might be better at discriminating certain antecedents. Despite Spector et al.'s (2006) compelling theoretical reasoning and findings for differential antecedents of the five-dimensional model, the dimensions suffer from low reliability with four out of the five dimensions having coefficient alphas below the recommended .70 (range = .55 to .85; *M* = .66; Nunnally, 1970). The authors argue that this is to be expected as their scale is formative. Despite this, the two-dimensional version of their measure has good internal consistency reliability (α = .86 for each dimension). Further, Spector et al. (2006) did not explicitly discuss validity evidence for their two or five-dimensional scale.

Eleven-dimensional Model. On a more extreme account, Gruys and Sackett (2003) divide CWB into eleven dimensions, including theft, destruction of property, misuse of information, misuse of time and resources, unsafe behavior, poor attendance, poor quality work, alcohol use, drug use, inappropriate verbal action, and inappropriate physical actions. Confirmatory factor analysis on their first study indicated that the 11-factor model had a moderate fit, while a one-factor model had a very poor fit. However, a follow-up principal component analysis and scree plot indicated that a single factor model provided the best and most parsimonious solution. A follow-up study using multidimensional scaling analysis indicated that the 11 categories could be categorized into two larger dimensions: an interpersonal-organizational dimension (similar to Robinson and Bennett's (1995) distinction) and a task relevance dimension. Together, the results are inconsistent and provided different dimensionalities of CWB, making it difficult to draw any solid conclusions. However, the emergence of the interpersonal-organizational dimension does provide further support for Robinson and Bennett's (1995) model.

Based on the current literature on the dimensionality of CWB, this study will consider a two-dimensional model of CWB (i.e., CWB directed towards the organization versus CWB directed towards people) in developing a new constructed response CWB SJT scale. The reason for going with the two-dimensional model are three-fold: (1) it has robust meta-analytic support of the differential antecedents between the dimensions (Berry et al., 2007), (2) it has better reliability and reported construct validity (Bennett & Robinson, 2000; Spector et al., 2006), and (3) has been argued to be the most frequently used model (Marcus et al., 2016; Berry et a., 2007).

Generic CWB Measures

Limitations. As CWBs are frowned upon and in some cases illegal, it is a sensitive construct to measure. Research has indicated that survey takers are less likely to disclose socially unacceptable behaviors (Tourangeau & Yan, 2007), which can result in nonresponse bias (Greco et al., 2015) or respondents indicating that they do not engage in CWBs when they do. In line with this, respondents might be wary of disclosing negative behaviors that they engage in due to the social desirability bias (Bennett & Robinson, 2000). To mitigate these issues, CWBs should be measured anonymously or confidentially to make the respondent more comfortable and likely to disclose their behaviors and reduce response biases (Bennett & Robinson, 2000). To continue, despite the limitations of self-reports of CWB, meta-analytic results have indicated that they result in very similar patterns of findings and even more reports of CWB compared to peer- and supervisor-reports and are, thus, an appropriate measurement method (Berry et al., 2012). Berry et al. (2012) describe that the reasons for this could include that peers and supervisors might not have the opportunity to observe all instances of CWB or discriminate more covert CWBs such as stealing or taking longer breaks than allowed.

Despite the popularity of the aforementioned CWB measures in the literature, Bowling and Gruys (2010) outline several important limitations of the measures that limit their use. More specifically, they report that CWBs are most commonly measured using generic measures (e.g., Bennet & Robinson 2000; Spector et al., 2006) that are assumed to be able to measure counterproductive behaviors for all types of jobs and organizations. They note that this is a faulty assumption as the type of counterproductive behavior employees can engage in can vary from job to job as well as from organization to organization. To this end, they argue that generic CWB measures include items that are not applicable across all jobs while excluding items that are important for many jobs. This can result in underestimation of scores on CWB when items are not applicable or when important counterproductive behaviors are missing, as employees will not have an opportunity to report their CWBs. For example, the item “Put in to be paid for more hours than worked” from the CWB measure by Spector et al. (2006) does not apply to salaried workers or those who do not put down the number of hours they work themselves. Similarly, the item “Dragged out work in order to get overtime” from Bennett and Robinson’s (2000) commonly used scale is not appropriate for employees who do not get overtime.

Approaches to Mitigate Limitations. To combat the issue of inapplicable items, researchers often exclude items they deem irrelevant (Greco et al., 2015). However, this in itself is problematic as removing or adding items to a validated scale can put into question the construct validity of the adapted scale (Heggestad, Schaef, Banks, Hausfeld, Tonidanel, Williams, 2019). This is certainly an issue as most research in the organizational sciences tend not to validate their adapted scales (Heggestad et al., 2019).

As another approach to mitigate the limitations of generic CWB measures, organizations can develop situation-specific CWB measures to ensure they are appropriate for their specific workplace and particular job in question (Bowling & Gruys, 2010). This method would allow for the incorporation of items that are important for specific jobs but are not captured by current scales. Examples of potential job-specific CWB items include falsifying data as a researcher, taking bribes as a politician (Bowling & Gruys, 2010), and unfair grading by teachers and professors. However, the approach of developing situation-specific measures has the limitations of being both time-consuming and costly, making it unlikely for researchers and organizations, especially smaller ones, to opt for this option over more accessible generic measures.

Another potential approach to mitigate the issues associated with generic CWB measures is to use an open-ended response format where employees can report the CWB they engage in without being restricted to the type of behaviors the question specifies. This would remove both the issue of unnecessary items being included in the measure, as well as important ones being omitted, as is the issue with current generic CWB measures. Further, this would also reduce the need for organizations having to develop CWB measures specific to a particular job at their organization.

Situational Judgment Tests

Traditional vs. Construct-specific SJTs

Traditional SJTs. Traditional SJTs have been used to assess judgment on work-related scenarios in order to measure job performance in various areas such as supervisory or managerial positions (Motowidlo et al., 1990; Wagner & Sternberg, 1991), team performance (Stevens & Campion, 1999), and conflict management (Olson-Buchanan et al., 1994). Factor analyses have revealed that traditional SJTs are multidimensional (Schmitt & Chan, 2006; McDaniel & Whetzel, 2005) and, as such, do not measure a single specific construct like what is usually the case with self-report rating, but have instead been found to correlate with different constructs (Schmitt & Chan, 2006; McDaniel & Whetzel, 2005). In a validation study of two SJTs by Weekley and Jones (1999) that included almost 4,000 employees across seven organizations, they found a significant weighted average correlation between the SJTs and job performance (.19), cognitive ability (.45), and job experience (.20). These findings have also been corroborated by meta-analytic results, which have found SJTs to be significantly correlated with cognitive ability ($p = .46$; McDaniel et al., 2001), personality dimensions of emotional stability, agreeableness, conscientiousness (mean $r = .31$, .26, and .25, respectively), and job experience ($r = .05$; McDaniel & Nguyen, 2001). Research has also found that traditional SJTs have incremental validity above and beyond cognitive ability, the Big 5, and a composite of the two in predicting job performance (McDaniel et al., 2007).

Construct-specific SJTs. More recently, construct-specific and unidimensional SJTs designed to measure specific constructs have been gaining increasing popularity (Guenole et al., 2017). This study will focus on creating an SJT specifically on the construct of CWB. Examples of construct-specific SJTs that have been previously developed include SJTs on employee integrity (Becker, 2005), team roles (Mumford et al., 2008), full-range leadership (Peus et al., 2013), personal initiative (Bledlow & Frese, 2009), and goal orientation (Westring et al., 2009). Research has shown that construct-driven SJTs can have high levels of validity. For example, Mussel et al. (2016) developed an SJT to measure each of the Big 5 dimensions (e.g., self-discipline and compliance) and found the average convergent and discriminant validity to be 0.59 and 0.01, respectively, while also finding support for the predictive validity. Similarly, in a development and validation study of a HEXACO SJT measure, Oostrom et al. (2019) found that the convergent and criterion-related validity was similar to or higher than traditional self-report measures. An added benefit of construct-driven SJTs over traditional SJTs is that they are more generalizable across jobs and organizations as they are less contextualized (Lievens, 2017). Overall, these findings support the utility of construct-driven SJTs as a valid assessment method in measuring specific constructs.

A contribution of this study is that while previous construct-specific SJTs have focused on measuring neutral or positive constructs and outcomes (e.g., integrity, team roles, personal initiative, goal orientation, and emotional intelligence), this study makes—to the knowledge of the author—a first attempt at using SJTs to measure negative workplace behavior (i.e., CWB). As respondents might differ in how they disclose negative behavior in this type of format, this study will shed some light on the appropriateness of using an SJT to measure negative workplace behavior.

Components of an SJT

The following section discusses the components that make up an SJT.

Situational Item Stems. Focal to any SJT are the situations (also known as item stems) that lay out the hypothetical scenario that the respondent must assess and respond to. These situations are usually developed using either critical incidents gathered from subject matter experts or through using a theoretical framework (Campion et al., 2014). As construct-specific SJTs, unlike traditional SJTs, are usually developed using a theoretical framework (Lievens, 2017), only this method will be discussed further here. This method involves the test developer drawing on literature and theory relevant to the construct of interest and using rational judgment to create situations (Campion et al., 2014). The aim of this method is to create scenarios that are relevant to the particular trait and thereby allow for its expression according to the trait activation theory (Tett & Burnett, 2003; Lievens, 2017), which holds that "… the behavioral expression of a trait requires arousal of that trait by trait-relevant situational cues" (Tett & Guterman, 2000, p. 398). The trait activation theory further holds that for optimal trait variance to be observed, the situation should contain a weak to moderate amount of situational information that is designed to elicit the relevant trait (Tett & Guterman, 2000; Snyder & Ickes, 1985; Weiss & Adler, 1984; Ekehammar, 1974). This is because too little trait-relevant situational information allows for few opportunities for the trait to be expressed, leading to little trait variance between those whose true scores are high versus low in the given trait; similarly, situations that are high in trait-relevant situational information might lead to increased trait expression by everyone, again leading to little trait variance between people (Tett & Guterman, 2000). Thus, situations that are weak to moderate in relevance to the trait are to be preferred for increased variance (Tett & Guterman, 2000; Snyder & Ickes, 1985; Weiss & Adler, 1984; Ekehammar, 1974).

Also relevant to item stem development is the response instructions. McDaniel, Whetzel, and Nguyen (2006) outlined two response instruction types common to SJTs: behavioral tendency and knowledge instructions. Behavioral tendency instructions tend to ask the respondent to indicate what they *would do* in a given situation, while knowledge instructions tend to ask what they think they *should do* (McDaniel & Whetzel, 2007). Research has indicated that behavioral tendency instructions tend to have higher correlations with personality traits and are thus more appropriate when measuring personality, while knowledge instructions tend to have a higher correlation with general cognitive ability and are more appropriate when measuring cognitive ability (McDaniel et al., 2007). Ployhart and Ehrhart (2003) further examined the differences between “Would do” and “Should do” response instructions on identical item stems and found that “Would do” instructions led to support of criterion-related validity, while this support was absent for “Should do” instructions. The authors also reported that including different response instruction types in an SJT can lower construct validity by making the test multidimensional. As such, it is favorable for construct-driven SJTs not to mix response instruction types. Due to behavioral tendency (i.e., “Would do”) instructions showing higher correlation with personality traits as well as leading to criterion-related validity, the current study will utilize “would do” response instructions for item stem development.

Response Formats. There are multiple different response formats available for SJTs, with the two big categories being forced-choice formats and constructed-response formats. In this study, constructed responses and open-ended responses will be used interchangeably to refer to the same response format. Forced-choice responses are most commonly used and include variations such as multiple-choice responses where the respondent picks the most appropriate answer, rating responses from most to least effective (or vice versa), rating effectiveness of each individual response, and ranking responses from most to least effective. The less commonly used constructed responses include formats such as open-ended written responses, oral responses, and video responses. Out of all the aforementioned formats, multiple-choice responses are most frequently used (Funke & Shuler, 1998).

Although multiple-choice formats are the most frequently used response method (Funke & Shuler, 1998), research has shown that open-ended response format SJTs—a format where respondents freely construct and input their response in a text-box—can have multiple benefits over multiple-choice responses. First, they can reduce racial adverse impact as shown by Lievens et al. (2019), who found that open-ended responses were associated with lower levels of differences between minority and majority groups, and by Edwards & Arthur (2007), who found reduced subgroup differences between African Americans and White SJT takers. This reduction in adverse impact has partially been attributed to decreased cognitive load (Lievens et al., 2019; Dahlke & Sackett, 2017) and increased test perceptions (Edwards & Arthur., 2007) associated with the constructed response format SJTs. Second, SJTs with open-ended responses have been indicated to have higher criterion-related validity (Funke & Schuler, 1998). Third, they have the potential for additional insight from respondents by not restricting them to fixed answers and allowing them to express ideas that might otherwise not have been part of the response options (Finch et al., 2018). Fourth, Kjell et al. (2019) argue that open-ended response formats mimic real-life communication and thinking more closely than closed-ended questions and, thus, have higher ecological and face validity. They further argue that it is the researcher’s responsibility to map open-ended responses into an interpretable scale rather than the participant's responsibility to translate their thoughts into a forced-choice scale. Fifth, drawing upon the theory of behavioral consistency (Wenimont & Campbell, 1968), which suggests that higher fidelity simulations and response options are better predictors of future performance than low fidelity simulation and response options, it could be argued that open-response SJTs are of higher fidelity than close-ended response options and might thus have better predictive validity of future performance. Open-ended responses are of higher fidelity as respondents have to come up with the best response themselves (Kjell et al., 2019) instead of selecting the best response from a pre-made list of options, which mimics real-life response situations more closely.

Scoring. However, despite its important benefits, open-ended response SJTs are rarely collected or analyzed (Edwards & Arthur 2007; Funke & Schuler, 1998). This might be attributed to the difficulty in scoring constructed response answers (Iliev et al., 2015). Previous research has almost exclusively utilized manual coding for analyzing open-ended response questions; a method associated with being time-consuming (Downer et al., 2019; Iliev et al., 2015; Roberts et al., 2014; Edwards & Arthur 2007), costly (Downer et al., 2019; Iliev et al., 2015; Roberts et al., 2014; Edwards & Arthur, 2007), labor-intensive (Esuli & Sebastiani, 2010), susceptible to possible reliability issues (Iliev et al., 2015; Lenhard et al., 2007), and rater effects (Lievens et al., 2019; Edwards & Arthur, 2007). To exemplify these limitations, Lievens et al. (2019) reported that it took raters approximately 35 minutes to score each respondent's open-ended SJT responses, while Funke & Schuler (1998) reported using three raters for rating each response to ensure interrater reliability. From this we can also deduce the high costs associated with the time and use of multiple raters. These limitations can quickly become increasingly problematic and infeasible as the number of SJT test takers increase (Iliev et al., 2015). However, research has indicated that these limitations can be mitigated with the use of computerized textual analysis—this will be discussed in a later section.

Some researchers have suggested future research to investigate computerized text analysis to analyze open-ended response SJTs (Lievens et al., 2019), while other researchers have recommended research into general constructed responses using text analysis (Iliev et al., 2015; Downer et al., 2019). To the author's knowledge, only Guo et al. (2021) have used computerized text analysis to analyze open-ended responses of an SJT. However, they did not develop the SJT themselves. It was instead developed by the Society of Industrial and Organizational Psychology (Thompson, 2019) as part of their yearly machine learning competition, and they have not published any paper or information on the development, reliability, or validity of the measure beyond its correlation with a multiple-choice personality scale.

With this in mind, and in response to calls by researchers, the current study will develop and validate an open-ended response format CWB SJT and analyze results using natural language processing to combat the limitations of manual coding while still benefiting from the strengths of open-response SJTs. Further, this will also allow for examining whether CWB can be appropriately measured in a constructed response format.

Validity and Reliability of SJTs

In construct-driven SJTs, construct validity is often assessed by examining the construct's correlation with the same or theoretically different constructs on self-report rating tests (Guenole et al., 2017). High correlation with the same or related construct and low correlation with a theoretically different construct would support convergent and discriminant validity, respectively.

To test reliability for construct-driven SJTs, Guenole et al. (2017) recommend examining internal consistency. While the authors do acknowledge that there is an inconsistency in the construct-driven SJT literature in terms of what type of reliability tends to be reported, they argue that internal consistency reliability is a necessary and appropriate reliability estimate for tests of unidimensional constructs where all items are intended to measure the same phenomena. Although some construct-driven SJTs have utilized test-retest and parallel forms reliability (e.g., Bledlow & Frese, 2009; Lievens & Sackett, 2007)—which are also the appropriate reliability estimates for traditional SJTs as they are multidimensional, making internal consistency reliability inappropriate—they should, if used, be used in conjunction with internal consistency reliability.

In terms of validity and reliability in the CWB literature, Bennet and Robinson's (2000) and Spector et al.'s (2006) scales will be discussed. Bennet and Robinson's (2000) measure examined convergent validity by assessing the correlation between their measure and similar constructs, such as production deviance and property deviance. They further assessed the association between their scale and theoretically relevant constructs, including frustration, procedural-, distributive-, and interactional justice, neglect, normlessness, Machiavellianism, and two facets of organizational citizenship behavior. To examine discriminant validity, the authors used responses to dissatisfaction, including exit, voice, and loyalty. Spector et al. (2006), on the other hand, did not explicitly discuss validity in their paper.

In terms of reliability, Both Bennett and Robinson (2000) and Spector et al. (2006) reported internal consistency estimates for each of their subscales. Bennet and Robinson's (2000) scale had a Cronbach's alpha of .81 for the organizational deviance dimension and .78 for the interpersonal deviance dimension. Spector et al.'s (2000) measure had comparable reliability, with CWB toward the organization having a reliability of .84 and CWB toward persons having .85. Informed by the discussed literature, the current study will examine the validity of the CWB

SJT using convergent and discriminant validity. Reliability will be assessed using test-retest reliability.

Computerized Text Analysis

Open-ended response questions are rarely utilized and reported compared to forced-choice measures (Roberts et al., 2014). When they are used, it is often for understanding ambiguities in less understood content areas and to identify opinions (Pietsch & Lessmann, 2018) rather than for construct measurement. However, with more recent developments in computerized text analysis and natural language processing (NLP), textual data has gained increasing popularity in the field of organizational sciences. Computerized text analysis refers to any type of automatic or semi-automatic analysis conducted by a computer on voluminous textual data with the intent to quantify, classify, predict phenomena, or discover trends (Gupta & Lehal, 2009; Harlow & Oswald, 2016). These methods can range in complexity from simple computer-aided text analysis (CATA)—which involves extracting patterns through word frequencies—to more advanced methods such as NLP, which also consider grammar, structure, and semantic meaning when analyzing text (Kobayashi et al., 2018a).

Benefits and Uses of Computerized Text Analysis

Computerized text analysis has multiple important benefits over manual coding. It has been indicated to reduce rater bias and error (Pietsch & Lessman, 2018; Roberts et al., 2014), be faster, more cost-effective (Roberts et al., 2014; Downer et al., 2019), and less labor-intensive than manual coding (Kobayashi et al., 2018a). These benefits are even further amplified with increased samples and text data (Campion et al., 2016). Further, Campion et al. (2016) showed that the use of computerized text analysis to score accomplishment records had comparable levels of validity and reliability as human raters scoring the same records and did not result in adverse impact.

Despite these benefits, the use of computerized text analysis in industrial and organizational psychology is still in its infancy; however, it has shown a wide range of uses. For example, Campion et al. (2016) demonstrated the utility of machine learning in scoring ~ 46,000 accomplishment records for selection purposes, Pandey and Pandey (2017) illustrated NLP as a way to help develop a measure of organizational culture using different text sources (e.g., letters to stakeholders), and Kobayashi et al. (2018b) showed how text classification could be used to extract job tasks from online job vacancy posts. However, little attention has been paid to leveraging computerized text analysis to analyze constructed survey responses in industrial and organizational psychology. This is surprising as constructed responses are a source of valuable and rich information (Esuli & Sebastiani, 2010).

Using Computerized Text Analysis for Construct Measurement

There are various ways of analyzing text data (e.g., topic modeling, sentiment analysis) depending on if one is trying to predict, cluster, or visualize data (Hickman et al., 2020). In other words, different techniques can be applied to the same text data to extract different types of information. Despite there being many different methods that can be used to analyze text data, they tend to follow the same overarching steps, including (1) preprocessing of the text data; this usually involves cleaning the text data and turning it into a format that can be analyzed further by a machine learning algorithm, and (2) implementation of the algorithm. When the aim is prediction, a last step of testing the model's performance and accuracy is also completed. In this section, we will discuss some of the various techniques that have been used to analyze constructed responses in the social science literature.

Supervised NLP. In order to score text data, supervised machine learning can be used. This involves providing a computer with correct scores for each text response to teach it how to emulate the scoring. After this is done, the model's accuracy is examined by testing it on a portion of data that the model has previously not seen. As an example, Guo et al. (2021) used NLP to analyze publicly available data on five open-ended responses to SJT questions; each item on the measure was designed to assess a Big Five personality dimension, with results being verified (i.e., showing convergent validity) by correlating them to numerical ratings on the Big Five scale. They used a Doc2Vec neural network model—a method that factors in context and word order to transform text data into numerical features at the document, sentence, and word level. These features were then used in a ridge regression to predict personality scores. This method was used to teach the model what numerical scores on the multiple-choice Big Five tend to go with what type of open-ended responses. Following this, the model performance was tested and validated on responses that the algorithm had previously not seen. They demonstrated that personality could be predicted from the constructed responses of an SJT with an average Pearson correlation of .28 (range = .22 to .38) with the multiple-choice measure of the Big Five. The authors additionally examined the performance of two other methods—*n*-gram bag-of-words (BOW) and linguistic inquiry and word count (LIWC)—that relax the requirement of an extremely large sample size (i.e., minimum of 10,000; Zhang et al., 2017) recommended for using Doc2Vec. Using these methods resulted in a .04 and .06 lower correlation between the constructed response scores and numerical scores for *n*-gram BOW and LIWC, respectively. This is notable as it highlights that simple methods that require smaller sample sizes had comparable, although slightly lower, performance compared to the more advanced Doc2Vec.

Topic Modeling. In contrast to using text analysis to quantitatively score constructed responses, other research has used topic modeling as a way to investigate trends in text data (e.g., Finch et al., 2018; Pietsch & Lessmann, 2018; Roberts et al., 2014). Topic modeling is a technique that involves a computer extracting latent topics or themes from a text based on word frequencies and co-occurrences and giving a statistical estimate of the presence of these topics in each text response (Kobayashi et al., 2018a; Blei et al., 2003; Finch et al., 2018).

Finch et al. (2018) utilized a topic modeling technique named latent Dirichlet allocation (LDA) to investigate themes present in open-ended survey responses and then used the resulting themes for further statistical analysis with close-ended responses. This study demonstrated that topic modeling is useful as it provides information about themes present in textual data while also allowing this information to be further used in combination with other data for statistical analysis. However, the authors argue that this method should be used as an additional tool to other methods (e.g., the resulting themes being used with close-ended responses for further analysis, or used in combination with manual coding of the open-ended responses) when analyzing open-ended responses, instead of solely by itself, in order to maximize utility.

In another study, Kjell et al. (2019) used latent semantic analysis (LSA)—a topic modeling technique that considers word patterns and the context that they are in to derive semantic similarity between words—to analyze open-ended survey responses designed to measure well-being, mental health problems, and evaluations of facial expressions. The results were then compared to numerical ratings of the same constructs. Their findings indicated that using LSA to analyze the open-ended responses yielded comparable or higher validity and reliability than the numerical rating scales of the same constructs.

Structural topic modeling (STM; Roberts et al., 2013; Roberts et al., 2014) is another topic modeling technique, which is unique in that it allows for the incorporation of a responses metadata (such as the responders' demographic information, time of response, etc.) for the topic modeling. The incorporation of metadata has been argued to make the STM more versatile, with potentially richer and more useful output (Roberts et al., 2014). For these reasons, STM has been described to be especially useful for open-ended survey responses (Roberts et al., 2014). As an example, Roberts et al. (2014) showed that in addition to finding topics present in documents, STM is able to indicate a topic's importance among different groups (e.g., by gender, political party, time point).

In sum, multiple different text analysis techniques have been applied to analyze constructed responses in the literature, some of which have been discussed here. The choice of which method to apply depends on what one is trying to achieve; for example, while topic modeling is usually used to categorize text into themes, other methods such as supervised machine learning are used to predict numerical scores. The current study will use supervised machine learning to score CWB responses and topic modeling to investigate themes present in responses.

Research Aims

As discussed, current CWB measures have important limitations that limit their utility and applicability for certain jobs. Using open-ended response format SJTs might relieve some of these limitations. In addition, constructed response SJTs have been indicated to have a number of advantages. Despite this, they are rarely used due to the difficulty and infeasibility associated with their manual coding. However, these issues could be mitigated with computerized text analysis. As such, the current paper (1) develops and validates a constructed response CWB SJT and (2) demonstrates the utility of leveraging computerized text analysis to measure CWB.

CHAPTER THREE: METHOD

This chapter is structured as follows: first, we discuss the sample. Next, we will go over the development and validation of the CWB SJT, followed by measures and procedures. Lastly, the data analyses will be discussed.

Sample

The participants were recruited from Mechanical Turk (MTurk)—a crowdsourcing website used to collect research data. Research has indicated the platform to yield comparable results and psychometric properties to traditional data collection methods, in addition to a more diverse and arguably generalizable sample (Buhrmester et al., 2011). A total of 776 participants took the first survey, out of which 530 passed at least two out of the three attention checks and were included for analysis. The final sample included full-time (i.e., 35+ hours/week) workers from diverse industries who were not self-employed. The majority were female (52.08%), white (78.11%), with a mean age of 43.15 (range 23 to 83; *SD* = 11.78), and mean tenure of 9.23 years (range 0.08 to 38.50; *SD* = 7.47). The sample size was above the minimum of 500 suggested for computerized text analysis (Ramineni & Williamson, 2013; Campion et al., 2016). Only employees with minimum 3-month tenure were included to make sure they have had opportunities to engage in CWB (Ciarlante, 2019). Self-employed employees were excluded as the likelihood and quality of the interpersonal and organizational CWB they engage in are likely to differ from other employed employees. A total of 503 participants took the second survey as well. Out of these, 467 passed at least two out of three attention checks and were kept for analysis.

Development of the CWB SJT

Item Development

A theoretical framework was used to develop the hypothetical scenarios (i.e., the item stems) of the CWB SJT. This involved drawing on relevant CWB literature and theory combined with using rational judgment (Campion et al., 2014). The aim was to develop scenarios that allow for the expression of CWB according to the trait activation theory (Tett & Guterman, 2000). This was achieved by developing scenarios that include trait-relevant situational cues. Weak to moderate amounts of trait-relevant situational cues were used to try to achieve optimal variance in CWB between the responses (Tett & Guterman, 2000; Snyder & Ickes, 1985; Weiss & Adler, 1984; Ekehammar, 1974). Further, all item stems included behavioral tendency response instructions that ask the respondents what they would do in the given scenario. This type of response instruction is appropriate as it is able to assess the behavioral tendencies of CWB.

There is little guidance in the extant literature for the number of SJT items to include for constructed response SJTs. However, the constructed response SJT of the Big Five developed by the Society of Industrial and Organizational Psychology (Thompson, 2019) consisted of one item per personality dimension that they measured. This current study took a more conservative approach by developing two SJT items for each CWB dimension (i.e., CWB-I and CWB-O) and required a minimum response length of fifty words for each item. This allowed for the sample of different behaviors and ensured the collection of ample text to be able to measure CWB-I and CWB-O. For the developed CWB SJT items, please see Appendix A.

Measurement Validation

Convergent validity was assessed by examining the correlations between the subscales of the CWB SJT (i.e., CWB-I and CWB-O) with the comparable subscales on the workplace deviance scale (Bennett & Robinson, 2000). Lower correlations between the opposite subscales on the two tests provide dimensionality evidence. As discussed, the workplace deviance scale was chosen as its conceptualization of a CWB into two dimensions of deviant behavior targeted toward the organization versus individuals has received a lot of support. Although Spector et al. (2006) also developed a CWB-C version that distinguishes between these dimensions, they used Bennett and Robinson's (2000) scale for their measurement development; as such, we opt for using the original scale by Bennett and Robinson (2000). As both the CWB SJT and workplace deviance measures are supposed to measure the same construct, scores on them should be positively and highly correlated. However, as the workplace deviance scale might underestimate actual CWB due to including items that might not be relevant to every occupation and excluding items that might be relevant to others (Bowling & Gruys, 2010), and due to the different formats and items between the SJT and the rating test, a moderate to high correlation is expected between the tests.

As another way to examine the construct validity of the CWB SJT, the correlation between the measure and the theoretically relevant constructs of procedural, distributive, and interactional justice (Colquitt, 2001) were examined. Distributive justice refers to how fair someone perceives their received outcomes (e.g., pay, promotions; Adams, 1965; Deutsch, 1985), while procedural justice refers to the perceived fairness of the procedures used to determine those outcomes (Leventhal, 1980). Interactional (or interpersonal) justice refers to the perceived quality of interpersonal treatment people receive in terms of the implementation of procedures and distribution of outcomes (Bies & Moag, 1986). Considerable research has linked perceptions of inequality and injustice with CWBs such as theft (Greenberg, 1990), vandalism (DeMore et al.,1988; Jermier et al., 1994), sabotage, withdrawal (Jermier et al., 1994), and retaliatory behavior (Skarlicki & Folger, 1997). These behavioral responses to injustice have been argued to be a way for employees to ameliorate their perceptions of injustice and inequality by "getting even" (Hollinger & Clark, 1983; Greenberg & Scott, 1996; DeMore et al., 1988). Based on this, organizational justice dimensions should be negatively associated with CWB-I and CWB-O.

To examine the discriminant validity of the CWB SJT scale, we used three ways employees respond to job dissatisfaction, namely through exit, voice, and loyalty (Hirschman, 1970). These constructs were also used by Bennett and Robinson (2000) for the discriminant validity of their workplace deviance scale. Rusbult et al. (1988) describe Hirschiman's (1970) original typologies as

> "*Exit* refers to leaving an organization by quitting, transferring, searching for a different job, or thinking about quitting. *Voice* describes actively and constructively trying to improve conditions through discussing problems with a supervisor or coworkers, taking action to solve problems, suggesting solutions, seeking help from an outside agency like a union, or whistle-blowing. *Loyalty* means passively but optimistically waiting for conditions to improve—giving public and private support to the organization, waiting and hoping for improvement, or practicing good citizenship." (p. 601).

From these definitions and in alignment with the predictions and results of Bennett and Robinson's (2000) study, it is expected that the subscales of the CWB SJT will have a low correlation with exit, voice, and loyalty.

The reliability of the CWB SJT was examined using test-retest reliability. In addition to the traditional measure of this reliability form, test-retest reliability was examined by examining the similarity of topics present in the CWB SJT from wave one to wave two using topic modeling.

Other Measures

The validity of the CWB SJT was assessed using the below constructs and measures.

Counterproductive Work Behavior

The Workplace Deviance scale (Bennett & Robinson, 2000) was used to measure CWB and its dimensions of organizational deviance (OD; equivalent to CWB-O) and interpersonal deviance (ID; equivalent to CWB-I). The scale includes nineteen items, with twelve items measuring OD and 7 items measuring ID. An example OD item is "*Spent too much time fantasizing or daydreaming instead of working,*" and an example ID item is "*Said something hurtful to someone at work.*" The items were measured on a Likert scale from 1 (*never*) to 5 (*always*). The internal consistency reliabilities were .89 for OD and .89 ID for the first wave of data collection.

Organizational Justice

Distributive justice (Colquitt, 2001; Leventhal, 1980) was measured using four items, with a sample item being "*Do your outcomes reflect the effort you have put into your work?*" Procedural justice (Colquitt, 2001; Thibaut & Walker, 1975; Leventhal, 1980) was measured using seven items, with an example item being "*Have you had influence over the outcomes arrived at by those procedures?*" Interpersonal justice was measured with a four-item scale, with a sample item being "*Have they treated you in a polite manner?*" All the justice scales were measured on a 5-point scale from 1 (*to a small extent*) to 5 (*to a large extent*). The scales had coefficient alphas of .95, .62, and .94, respectively.

Exit, Voice, and Loyalty

Exit and loyalty (Farrell, 1983) were measured with three items each, while voice was measured using an adaption of the four-item voice scale by Rusbult et al. (1988). All items were measured on a Likert scale from 1 (*never*) to 5 (*always*). Sample items include "*Getting into action and looking for another job*" for exit, "*Waiting patiently and hoping the problem will solve itself*" for loyalty, and "*Going to my immediate supervisor to discuss the problem*" for voice. The Cronbach's alpha for these scales were .81, .62, and .64, respectively.

Social Desirability

In line with multiple previous studies on faking (e.g., Ones et al., 1996), social desirability (Hays et al., 1989) will be measured as an indicator of potential faking on the CWB SJT. This construct was measured with five items, with a sample item being "*I am always courteous even to people who are disagreeable*." The items were measured on a Likert scale from 1 (*strongly disagree*) to 5 (*strongly agree*). The measure had a Cronbach's alpha of .78.

Neuroticism

In order to examine whether the CWB SJT measures CWB tendencies rather than CWBs, neuroticism, trait self-control, and problem-focused coping were also measured. The neuroticism (Donnellan et al., 2006) scale included four items, with an example item being "*I have frequent* *mood swings*." The items were measured on a Likert scale from 1 (*strongly disagree*) to 5 (*strongly agree*). The internal consistency reliability of the scale was .81.

Trait Self-control

Trait self-control was measured using the thirteen-item scale by Tagney et al. (2004). A sample item from this scale is *"I am good at resisting temptation."* The items were measured on a Likert scale from 1 (*strongly disagree*) to 5 (*strongly agree*). The Cronbach's alpha of this scale was .91.

Problem-focused Coping

Problem-focused coping was measured using the problem-focused coping subscale from Carver et al.'s (1989) coping measure. This subscale includes four items, with an example item being *"I try to come up with a strategy about what to do."* The items were measured on a Likert scale from 1 (*strongly disagree*) to 5 (*strongly agree*). This measure had a Cronbach's alpha of .73.

Procedure

Data were collected in two waves through the distribution of online surveys on MTurk, with one month between each wave. In the first wave, the CWB SJT, workplace deviance, organizational justice (i.e., including distributive, procedural, and interpersonal justice scales), exit, voice, loyalty, social desirability, neuroticism, trait self-control, problem-focused coping, and demographic and job-related questions were included. In the second wave, the CWB SJT was measured to for test-rested reliability. Attention checks were included in both waves. Participants were compensated $1.50 for completing the first survey and $1.25 for the second.

Data Analysis

The analyses included (1) preprocessing the textual data, (2) implementing the machine learning algorithm, and (3) testing the performance and accuracy of the model. All analyses were conducted using *R* version 4.1.1 (R Core Team, 2021).

Data Preprocessing

Text preprocessing involves cleaning and transforming text data into a format that can be more easily and accurately analyzed (Kobayashi et al., 2018a). To start, all the responses within each CWB dimension were merged for each respondent. Next, Hickman et al.'s (2020) recommendations for preprocessing text were followed. First, spelling was corrected, followed by converting all text to lowercase so the computer would not see words with capitalizations as different from the same words without capitalization (e.g., *Happy* and *happy*). Next, contractions were expanded. Then, lemmatization was applied. This involved removing inflections from words to turn them into their base form (e.g., happiest and happily become happy), allowing the same words with different inflections to be grouped together for analysis. Following this, handle negation was applied by adding an underscore (i.e., "_") after negation words. This allowed the computer to be able to differentiate between, for example, "*happy*" and "*not happy*." Lastly, stop words, non-alphabetic characters (i.e., numbers, punctuation, and symbols), and extra white space were removed. Stop words refer to common words such as "the," "is," and "are." These steps were completed with the *tm* package (Feinerer, 2018; Feinerer et al., 2008) for lower case conversion and removal of non-alphabetic characters and white space; the *tm* and *stopwords* (Benoit et al., 2021) package for removal of stop words; *R*'s base packages for handle negation; the *hunspell* (Ooms, 2018) package for correcting spelling; *qdap* (Rinker et al., 2020) for expanding contractions; and the *textstem* (Rinker, 2018) package for lemmatization. Together, these steps made the vocabulary smaller by making the word formatting more consistent across the corpus (i.e., all the text data), overall helping increase the validity and power of subsequent NLP analyses (Hickman et al., 2020).

After the corpus was cleaned, it was transformed to a format that allowed for further analysis. The first step in this process was to tokenize the corpus into *n*-grams. *n*-grams are a contiguous sequence of *n* words, with commonly used ones being unigrams (i.e., consisting of a single word), bigrams (i.e., consisting of two words), and trigrams (i.e., consisting of three words). Hickman et al. (2020) argue that bigrams and trigrams can have added validity over unigrams due to incorporating semantic information, making them potentially more useful when analyzing constructed survey responses. However, larger *n*-grams have the caveat of running into sparsity issues, especially when the corpus is small. In light of this, the current study tokenized the text data into unigrams and bigrams using *text2vec* (Selivanov et al., 2020).

Very frequent (i.e., occurring in over 99% of the documents) and infrequent (i.e., occurring in less than 2% of the documents) terms were removed from the vocabulary as these words were likely to provide little discriminant power in subsequent analyses (Kobayashi et al., 2018a). Next, the corpus was transformed to a document-term matrix format where each row designates a document (i.e., the merged constructed responses for either CWB-O or CWB-I), columns designate *n*-grams, and cells designate *n*-gram frequencies in each document. Two document-term matrices were created, one for CWB-O and one for CWB-I. This was accomplished using the *tm* package (Feinerer, 2018). For more details about each of these preprocessing steps, please see Hickman et al. (2020).

Machine Learning

All the steps in this section were completed twice, once for creating a model for CWB-O and once for CWB-I. Following the text preprocessing and data transformation, the data was divided into a train and test data set using a 90%/10% split. The train data was used to create the model, while the test data was used to assess the model's accuracy on data it had not been trained on. The next step involved implementing an appropriate supervised machine learning algorithm and train the model. In the context of this study, this process involved using the numerical ratings of CWB-O and CWB-I from the workplace deviance scale (Bennett & Robinson, 2000) to train the computer on how to score constructed responses of the CWB SJT for the two CWB dimensions. The model was trained using least absolute shrinkage and selection operator (lasso) regression (Tibshirani, 1996). This method was also used by Guo et al. (2021) to predict company ratings from company reviews. This method is able to shrink less important coefficients to zero, thereby reducing the number of features and complexity of models (Putka et al., 2018). As a shrinkage method, lasso regression is especially useful when there are many features (Friedman et al., 2010), something that is usually the case when dealing with text data. After running the regression, predicted numerical scores are outputted; in the context of this study, this output will be each respondent’s measurement of CWB-I and CWB-O. Please see Tibshirani (1996) and Putka et al. (2018) for further information about lasso regression.

The *caret* (Kuhn, 2020) package was used with a five-fold cross-validation with three repetitions to implement lasso regression and train the model. Cross-validation helps estimate and minimize overfitting and, thus, improve model generalizability (Hastie et al., 2009). The tuning parameters used can be found in Appendix B. Following this, the performance of the created models (i.e., one model for CWB-I and one for CWB-O) were tested on the test data sets without labels (i.e., the CWB dimension scores) present by examining the correlation between the predicted values and actual values on the test data set. The model performance was further assessed by examining the resulting root mean square error (RMSE), mean absolute error (MAE), and *R*-square (R^2) values on the test data sets. For the full *R* code that was used, please see Appendix B.

Topic Modeling

Roberts et al. (2013; Roberts et al., 2014) have proposed a new topic modeling method, STM, which has been described to be especially useful for analyzing open-ended survey responses by making the analysis easier and more revealing. STM is a probabilistic modeling method that extracts topics in documents or textual responses using unsupervised machine learning and natural language processing. Through this process, topics are created through the mixture of words available in the corpus. In turn, each document or constructed response is assigned a mix of topics of varying proportions based on the words available in the document. In other words, as a mixed-membership model, each document can be assigned one or more topics, while each term in a document belongs to a single topic. This typically results in documents that are composed of different topics in different proportions. The current research uses STM as an exploratory way to examine themes present in the constructed CWB SJT responses. The themes should be relevant to the STJ question while also being mostly consistent from one time point to the next.

The researcher has to set the number of topics for STM, similar to what is done in exploratory factor analysis. To find the optimal number of topics for the model, a combination of (1) trying different number of topics, (2) content review of topics and words, and (3) examining the semantic coherence (i.e., the level co-occurrence of words within a topic), exclusivity (i.e., the extent words within a topic do not occur in other topics; Roberts et al., 2014), hold-out likelihood (i.e., probability of held-out documents based on training documents; Wallach et al., 2009), lower bound (i.e., an examination of convergence by assessing change in variational lower bound; Roberts et al., 2019), and residuals was used. The *stm* package (Roberts et al., 2019) is used to find the number of topics that maximize these statistics. Next, the *stm* (Roberts et al., 2019) package was used to conduct the topic modeling on the CWB-I and CWB-O responses. For the *R* code used for the STM, please see Appendix C.

CHAPTER FOUR: RESULTS

Machine Learning

Construct Validity

RMSE, MAE, R^2, and the correlations between predicted and actual values on the test data set are reported in Table 1. We found that the RMSE and MAE were small, which indicates an acceptable model. Further, the correlations between the machine learning scored CWB SJT responses and the numerical ratings of CWB-O ($r = .54$, $p < .05$) and CWB-I ($r = .48$, $p < .05$) scores were moderate to strong, like predicted. The SJT responses prediction of CWB-O ($R^2 = .29$) and CWB-I ($R^2 = .23$) were comparable. These findings indicated that constructed response of the CWB SJT scored by the machine learning model could predict numerical ratings of CWB to a moderate degree, suggesting convergent validity.

Table 1

Evaluation of the Lasso Regression Model's Performance on the Test Dataset

CWB Dimension	RMSE	MAE	R^2	r
CWB-O	.48	.35	.29	.54*
CWB-I	.57	.39	.23	.48*

Note. RMSE = root mean square error, MAE = mean absolute error.
* $p < .05$, ** $p < .01$.

Consistent with predictions, exit, voice, and loyalty all had small and insignificant correlations with CWB-O and CWB-I as measured with the CWB SJT (see Table 2), indicating discriminant validity. The correlations between CWB SJT dimensions and the theoretically relevant constructs of procedural, distributive, and interpersonal justice were smaller than expected and insignificant for both CWB-O and CWB-I. The correlation between social desirability and CWB-O (*r* = .28, *p* < .05) was positively significant, while the correlation with CWB-I (*r* = .26, *p* > .05) was not. Similarly, the correlation between neuroticism and CWB-O (*r* = .27, *p* < .05) was positively significant, while the correlation with CWB-I (*r* = .19, *p* > .05) was insignificant. Trait self-control was significantly negatively correlated with both CWB-O (*r* = -.27, *p* <.05) and CWB-I (*r* = -.32, *p* <.05). Problem-focused coping was not correlated with either CWB-O (*r* = -.18, *p* >.05) or CWB-I (*r* = .00, *p* >.05). As the test data used to examine the discussed correlations was small at 52, it might explain why some of the moderate correlations are insignificant despite their correlation size.

The correlations between the discussed constructs and CWB dimensions from the workplace deviance scale (Bennett & Robinson, 2000) were also examined (see Table 2). Findings from this indicate that the workplace deviance scale dimensions, overall, had significant correlations with the justice measures and with exit and loyalty. Further, CWB-O from the workplace deviance scale was significantly correlated with social desirability (*r* = .17, *p* < .01), trait self-control (*r* = -.48, *p* < .01), and problem-focused coping (*r* = -.24, *p* < .01). Similarly, CWB-I was significantly correlated with social desirability (*r* = .19, *p* < .01), trait self-control (*r* = -.33, *p* < .01), and problem-focused coping (*r* = -.16, *p* < .01). Neither CWB-O (*r* = -.06, *p* > .05) nor CWB-I from the workplace deviance scale were significantly related to neuroticism (*r* = .04, *p* > .05).

The correlation between CWB-O and CWB-I on the SJT was .55 (*p* < .01), while the corresponding correlation on the workplace deviance scale was .63 (*p* < .01). This provides dimensionality evidence for the CWB SJT, as the test is able to distinguish the CWB dimensions.

Table 2

Correlations between CWB and other constructs

Comparison measure	Correlations			
	CWB SJT		Workplace Deviance Scale	
	CWB-O	CWB-I	CWB-O	CWB-I
Theoretically related behaviors				
Procedural justice (Colquitt, 2021; Thibaut & Walker, 1975; Leventhal, 1980)	.18	-.10	-.07	-.10*
Distributive justice (Colquitt, 2021; Leventhal, 1980)	-.26	-.11	-.25**	-.11*
Interpersonal justice (Colquitt, 2021; Bies & Moag, 1986)	-.07	-.03	-.32**	-.23**
Dissimilar behaviors				
Exit (Farrell, 1983)	.05	.11	.30**	.18**
Voice (Rusbult et al., 1988)	-.01	.07	.04	.08
Loyalty (Farrell, 1983)	-.17	-.05	.15**	.10*
Other behaviors and characteristics				
Social desirability (Hays et al., 1989)	.28*	.26	.17**	.19**
Neuroticism (Donnellan et al., 2006)	.27*	.19	-.06	.04
Trait self-control (Tagney et al., 2004)	-.27*	-.32*	-.48**	-.33**
Problem-focused coping (Carver et al., 1989)	-.18	.00	-.24**	-.16**

Note. *N* = 52 (i.e., the test data set) for the CWB SJT correlations; *N* = 530 for the workplace deviance scale correlations. All scales were measured at wave one. * *p* <.05, ** *p* <.01.

Test-retest Reliability

Test-retest reliability was examined by looking at the correlation between CWB-O and CWB-I SJT scores from wave one with corresponding scores on wave two. Wave two data were scored by applying the created machine learning prediction model to it. The findings indicated no test-retest reliability of CWB-O (*r* = -.09, *p* > .05) or CWB-I (*r* = -.19, *p* > .05). A supplemental analysis was run to examine the test-retest reliability by creating the machine learning prediction model on the entirety of wave one data (n = 530) instead of only the train data portion (n = 478) of it, as this allowed for training the model on a slightly larger data set. However, results from this also indicated no test-retest reliability.

Topic Modeling

Structural topic modeling was used as an exploratory way to further examine trends in the data. Topics with sizes two through twenty were examined to find the ideal number of topics to model. Semantic coherence and exclusivity scores were maximized for models with 4, 5, and 6 topics, indicating that these topic sizes might be ideal (see Figure 1). As an additional measure of examining the ideal number of topics, individual metrics for semantic coherence, held-out likelihood, lower bound, and residuals for each model were also examined (see Figure 2), as recommended by Roberts et al. (2019). To find the optimal topic number according to these metrics, the topic model should have as high semantic coherence, held-out likelihood, and lower bound scores as possible while minimizing residuals (Roberts et al., 2019). The findings suggested that five topics might be most appropriate. Finally, the author examined the topics of the models with four, five, and six topics by examining the words within each topic. Findings from this also indicated five topics to be the best as the topics made the most sense by being clear and distinguished from each other. Please see Table 3 for the top ten words represented under each of these topics as judged by frequency and exclusivity (FREX) scores, which are frequently used to estimate top words in topics when using STM (Roberts et al., 2014). The first topic was about looking for a new job, the second topic included terms related to communication, the third topic contained words related to project completion and communication, the fourth topic was somewhat ambiguous but was loosely related to hard work, and the fifth topic centered around words related to the first CWB SJT item.

The change in topics from wave one to wave two responses was also examined. Findings indicated that the topics were largely represented to the same extent in both waves (see Figure 3). This is expected as the SJT items asked the same question in both waves; thus, the topics in both waves should be represented similarly. This finding also suggests the consistency of topics over time, which provides general test-retest reliability evidence for the CWB SJT. A similar pattern of results—in terms of topic representation from wave one to wave two—was found for the individual CWB SJT items; for full topic modeling results for each individual CWB SJT item, please see Appendix D.

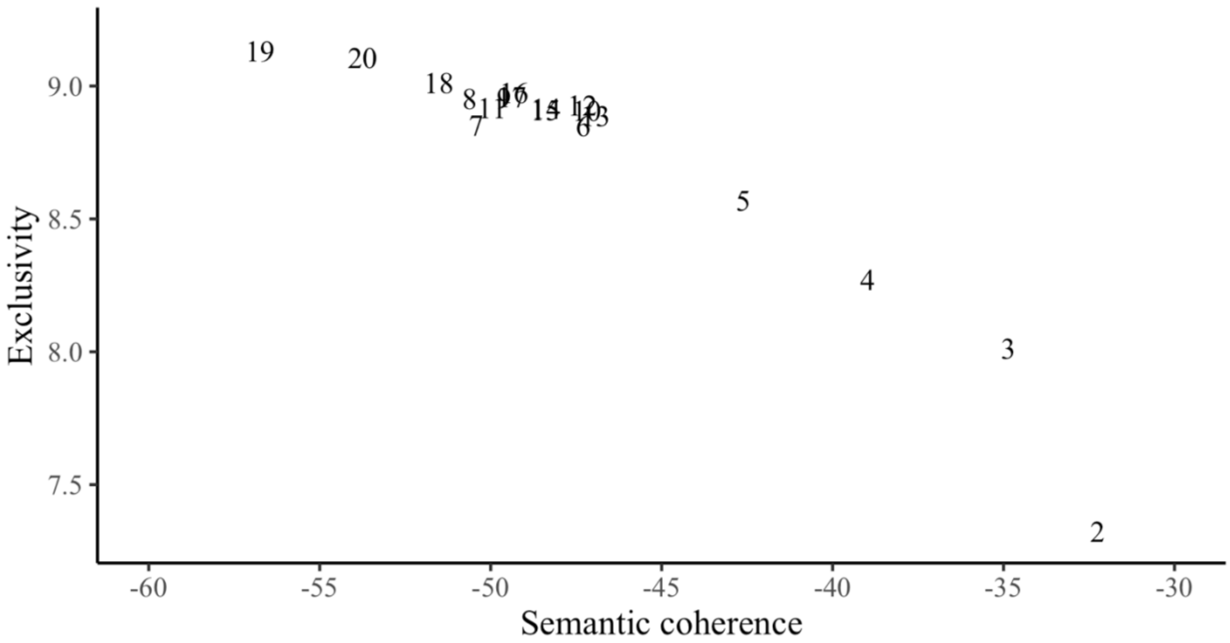


Figure 1
Semantic Coherence and Exclusivity Scores for Topics 2 through 20

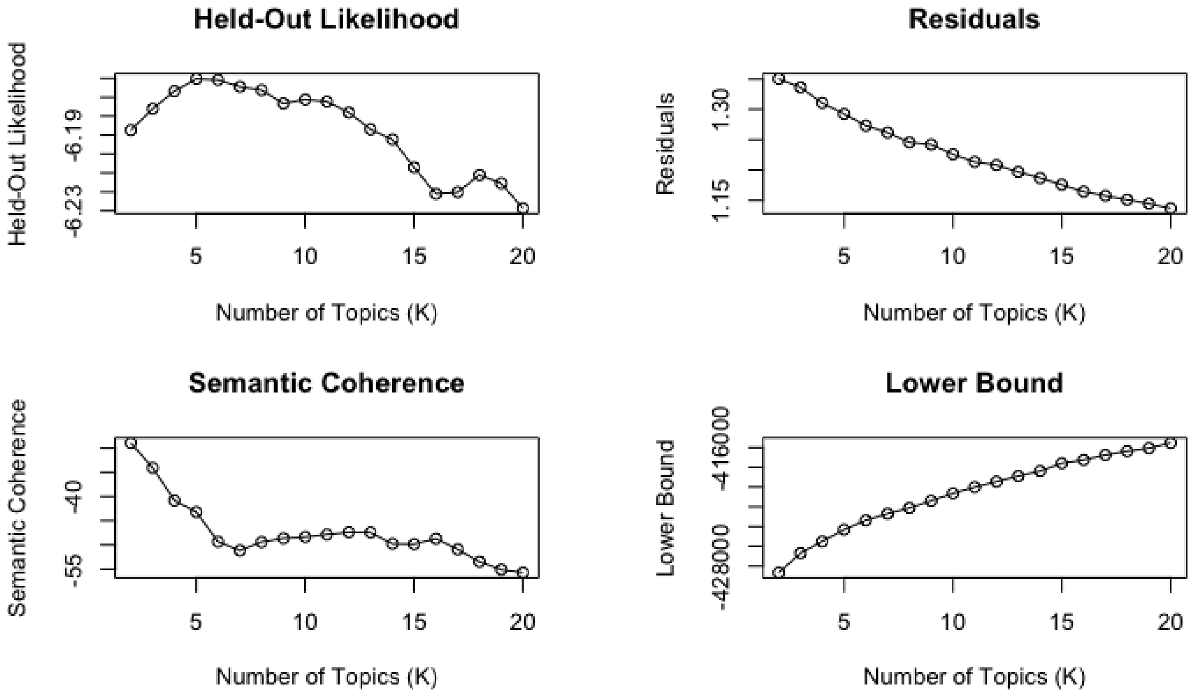


Figure 2
Evaluation Metrics of Topic Models with Topics Ranging from 2 to 20

Table 3

Topics and Top Words Representing Them According to FREX Scores.

Topics				
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
another_job	ask_raise	meet	just	estimate
start_look	explain	discuss	get_pay	earnings
look_another	ask	meet_supervisor	anything	datum
look_new	supervisor_ask	complete_project	hard	potential
look	supervisor_explain	deadline	work_hard	base
new_job	extension	meet_deadline	else	quarterly_earnings
start	load	complete_task	everyone	quarterly
another	_ask	review	like	present
job	ask_supervisor	task	slow	note
elsewhere	talk_supervisor	hour	think	new_client

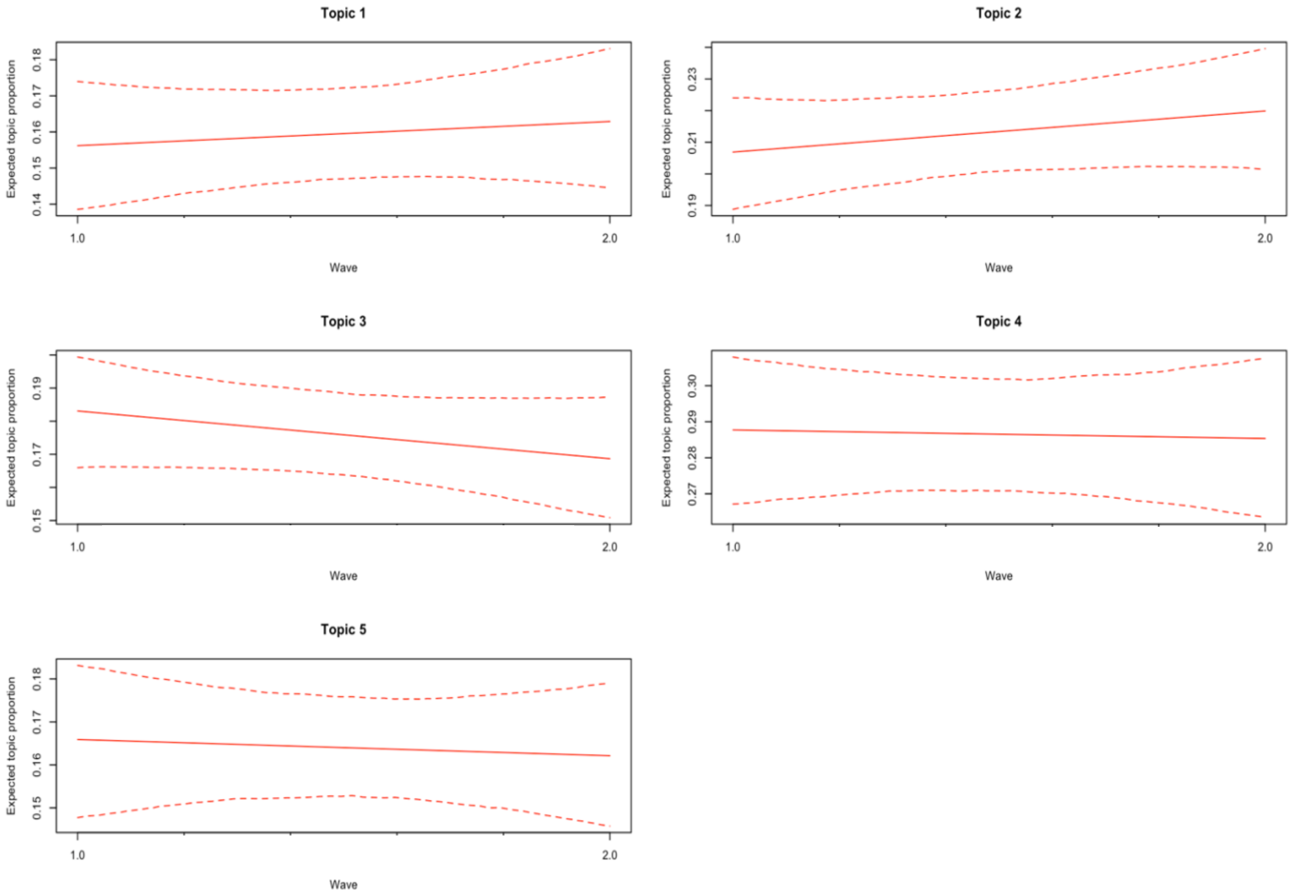


Figure 3

Change in Topic Prevalence from Wave 1 to Wave 2

CHAPTER FIVE: DISCUSSION

There are some limitations inherent to current CWB measures that are somewhat problematic, including that they contain items that do not apply to all jobs while missing items that are important for other jobs (Bowling & Gruys, 2010). The current study responded to these issues by drawing on the benefits associated with constructed response SJTs, such as them having potential for more insight from respondents (Finch et al., 2018), to develop an open-ended response CWB SJT. To minimize the drawbacks currently associated with the manual analysis of open-ended response SJTs (e.g., being time-consuming and costly)—which is also a reason why they are rarely used—the study leveraged natural language processing in combination with machine learning to semi-automatically measure CWB-O and CWB-I.

The findings indicated that the machine learning model was able to predict numerical CWB-O and CWB-I scores (Bennett & Robinson, 2000) from the constructed responses of the CWB SJT to a moderate to strong degree. The CWB SJT did not seem to correlate with the predicted theoretically similar behaviors of procedural, distributive, and interactional justice; however, it was related to variables that are more indicative of CWB tendencies, specifically neuroticism and trait self-control (Kozako et al., 2013; de Boer et al., 2015). This suggests that the CWB SJT might be better used as an alternative or as a supplementary method of approximating measurement of CWB tendency rather than enacted CWB. The discrepancy between the CWB SJT and the workplace deviance scale in terms of correlating with current organizational perceptions (i.e., organizational justice) highlights this. From this perspective, the lack of correlation between the CWB SJT and the justice measures makes sense, as current organizational perceptions should be irrelevant when answering hypothetical SJT scenarios. In further support of this view, while CWB-O from the CWB SJT correlated with neuroticism, none of the workplace deviance scale dimensions did. This further indicates that the CWB SJT might better capture CWB tendencies than enacted CWB.

Findings also suggested that CWB-O was significantly related to social desirability. According to some previous literature, this indicates that individuals might have been engaging in faking while completing the CWB SJT (e.g., Ones et al., 1996, Hough et al., 1990). However, according to some research, social desirability is not a good indicator of faking (Peterson et al., 2011), and its use to identify or correct faking is questionable (Burns & Christiansen, 2006). It has also been questioned whether social desirability instead measures personality variance or individual differences in responding styles (Smith & Ellingson, 2002). For example, social desirability has been found to be significantly related to both conscientiousness and neuroticism (Smith & Ellingson, 2002). Due to the conflicting literature on this, while the CWB-O is correlated with social desirability, the potential implications of this on faking are not entirely clear.

A limitation is that the CWB SJT had no test-retest reliability based on correlation. This might be due to the prediction model overfitting on the train and test data. In some situations, a separate validation data set can be used to examine overfitting, but since the sample size was very small in this study, this was not done. Additionally, it has been argued that one is especially likely to run into overfitting issues when there are a lot of predictors in the data (Chen & Wojcik, 2016), something that tends to be the case with text data and is the case with the current study which had 653 unigrams and bigrams as predictors. This is an especially big issue with models that rely on small data sets. Additionally, it might be that people responded qualitatively differently when they retook the CWB SJT due to already being familiar with the situational scenarios and having responded to them. It could be that another type of reliability, such as internal consistency reliability, might be a more appropriate measure of reliability for the CWB SJT. Although this type of reliability was not examined in this study due to the small sample size, it certainly would be an interesting thing to assess. Nevertheless, as an alternative measure of test-retest reliability, the consistency of topics from wave one to wave two was examined using topic modeling, and the findings suggested that topics were consistent across waves; in other words, topics that were discussed were discussed the same extent in both data collection waves. While this does not replace traditional test-retest reliability, it does provide some evidence that, on the whole, when measured at two different time points, topics that emerge have high consistency over time.

Theoretical and Practical Implications

Current CWB measures face the issue of including items that are irrelevant in certain jobs while excluding items that are important for others (Bowling & Gruys, 2010). This study extends this CWB literature by creating an SJT that measures CWB tendency, and that is not occupation-specific. The developed constructed response SJT allows respondents to freely write out their answers without being constrained to response options predetermined by the researcher, therefore also allowing for potential insight into how respondents would think or approach a difficult situation. The study also extends the literature on SJTs by making a first attempt at using an SJT to measure negative behavior, as previous construct-specific SJTs have focused on measuring neutral or positive constructs (e.g., goal orientation, emotional intelligence, and integrity; Westring et al., 2009; Sharma et al., 2013; Becker, 2005).

An important practical implication of this study is that it provides evidence that machine learning could be used to score open-ended response SJTs on a specific construct. With this, organizations can draw on the benefits of constructed response SJTs, while limiting their drawbacks. In other words, computerized analysis helps make open-ended response SJT more accessible and feasible to organizations by reducing time, costs, and labor associated with their manual analysis. Similarly, the study provides a stepping-stone for researchers interested in drawing on the benefits of constructed response SJTs but are hesitant due to the difficulty in analyzing the responses. This is especially important as open-ended responses can hold valuable information that is otherwise not always captured by forced-choice options. A further benefit is that free statistical software programs, such as *R* (R Core Team, 2021), can be used to analyze text data. R code is provided (see Appendix B and C) so that even people with little background in text analysis can use the measure. This code can also be adapted to train and develop models for other open-ended response measures.

Limitations and Future Directions

There are a few limitations to the current research. Due to the analytical methods used, the test data set—which the reliability and validity were examined on—was small. This caused the assessment of reliability and construct validity through correlations to be underpowered. A larger sample size could also further improve the prediction and accuracy of the machine learning model. Thus, future research should replicate this study with a larger sample size to create a more robust model for the CWB SJT. Another limitation of the current study is that a simple *n*-gram approach was used for the NLP instead of a more advanced method like Doc2Vec. However, research has indicated that advanced techniques do not always perform better (e.g., Guo et al., 2021). Nevertheless, future research can investigate the performance of an advanced NLP approach in analyzing responses of the CWB SJT to see if there are any performance improvements. A further limitation of the study is that the participants were recruited from MTurk instead of an actual organization. However, research has indicated that this platform has comparable results to traditional data collection methods (Buhrmester et al., 2011). Further, despite the survey being designed to be anonymous, distributing it through MTurk arguably gives respondents an extra sense of anonymity compared to distributing it in an actual organization. This could make respondents feel more comfortable in disclosing their potential negative behaviors. Additionally, as the CWB SJT can be used regardless of one's occupation—which is a distinguishing feature of this CWB measure compared to previous measures—MTurk's mixed employee pool is a suitable source for recruiting participants. A further limitation is that constructed responses, by nature, take longer to fill out compared to forced-choice formats (Kjell et al., 2019). This is something the researcher must consider as longer surveys can result in lower completion rates (Liu & Wronski, 2017). However, constructed responses do have the advantage of, for example, better criterion-related validity (Funke & Schuler, 1998), ecological validity, and fidelity (Kjell et al., 2019), as well as lower adverse impact (Lievens et al., 2019).

There are multiple interesting future research directions to the current study. Future research should further scrutinize the developed CWB SJT. One important way this can be done is by analyzing the CWB SJT responses with both manual coding and computerized text analysis to test and compare results. If the manual coding scores are comparable to the computerized scoring, it provides evidence for the utility of the computerized approach. To further test the validity of the developed measure, other constructs such as personality variables (e.g., negative affect, trait anger) or stressors (e.g., organizational constraints) could be used as predictors of CWB. If correlations are found, this will provide validity evidence for the CWB SJT. One way future research can potentially combat the social desirability that was found to be associated with the CWB SJT is by adapting the current SJT items to refer to coworkers instead of the respondent themself. This can be done by adjusting the existing items from “what would you do?” to “what would your coworker do?”

Future research should also consider developing other open-ended response SJTs and use computerized text analysis to draw on the befits of constructed response SJTs. While the current study developed a construct-specific SJT, future research can also apply discussed methods to traditional SJTs that tend to be multidimensional (Schmitt & Chan, 2006; McDaniel & Whetzel, 2005). Future research should also examine other types of data, such as video and audio responses, which Lievens et al. (2019) found to have added advantages over written responses, such as less adverse impact. These responses can then be transcribed into text format by computer programs and then analyzed through similar methods described in this paper.

Conclusions

The current study developed and validated a new CWB SJT measure that utilizes an open-ended response format. Through this, the study addressed limitations associated with current CWB measures, such as containing items that do not apply to all jobs while missing items that are important for other jobs. Further, the study drew on the benefits of open-ended response SJTs to create a CWB tendency measure while simultaneously extending the SJT literature by demonstrating that computerized text analysis can be used as a less time-consuming, more cost-efficient (Downer et al., 2019; Iliev et al., 2015), and less labor-intensive (Esuli & Sebastiani, 2010) way of analyzing open-ended SJT responses.

APPENDIX A: MEASURES

Inclusion Criteria

- Please enter your MTurk ID below
  - Open-ended response
- Are you over the age of 21?
  - Yes
  - No*
- How long have you been employed at your current position?
  - Less than 3 months*
  - More than 3 months
- On average, how many hours do you work per week?
  - Less than 20 hours*
  - 20-34 hours*
  - 35 or more hours
- Do you interact with coworker's and/or supervisors on a weekly basis?
  - Yes
  - Sometimes
  - No*
- What percentage of your worktime do you spend working in teams?
  - 0-100% scale**
- Are you self-employed?
  - Yes*
  - No
- Do you reside in the United States?
  - Yes
  - No*

*Subject were screened-out

** If answers were less than 10%, subject were screened out

CWB Situational Judgment Test

Please read the following scenario and respond as thoroughly as possible.

- You have been asked to calculate and deliver your company's quarterly earnings to a potential new client. The deadline is tomorrow and there is no way you can finish the job in time. You have calculated some of the metrics and believe you might be able to guess the rest from your current calculations in order to turn something in. Successfully completing the project on time would result in a substantial performance bonus. What would you do and why?
- You consistently outperform your coworkers, causing your supervisor to give you more work. Despite doing more work, you receive the same pay and recognition from your supervisor. You feel like your work is not being appreciated. What would you do and why?
- You have noticed one of your coworkers consistently making rude remarks about you. You have tried to talk to the coworker but they have not stopped. Your supervisor tasks you and your rude coworker with an important project that needs to be finished by tomorrow morning. Due to the amount of work, you and your coworker continue working

after everyone else has left. During this time, your coworker continues mistreating you (for example, by yelling). What would you do and why?

- You have not gotten a raise in a long time, and you think you deserve it. When asked, your supervisor said that the company cannot afford raises. You learn that a new and less experienced employee in the same position and department as you earns more money than you. Your supervisor tasks you and the new employee with an important project and asks you to lead since you have more experience. How would you handle this project and why?

Open-ended response

Workplace Deviance (Bennett & Robinson, 2000)

To what extent have you engaged in each of these behaviors in the last year:

- Organizational Deviance
    1. Taken property from work without permission
    2. Spent too much time fantasizing or daydreaming instead of working
    3. Falsified a receipt to get reimbursed for more money than you spent on business expenses
    4. Taken an additional or longer break than is acceptable at your workplace
    5. Come in late to work without permission
    6. Littered your work environment
    7. Neglected to follow your boss's instructions
    8. Intentionally worked slower than you could have worked
    9. Discussed confidential company information with an unauthorized person
    10. Used an illegal drug or consumed alcohol on the job
    11. Put little effort into your work
    12. Dragged out work in order to get overtime
- Interpersonal Deviance
    1. Made fun of someone at work
    2. Said something hurtful to someone at work
    3. Made an ethnic, religious, or racial remark at work
    4. Cursed at someone at work
    5. Played a mean prank on someone at work
    6. Acted rudely toward someone at work
    7. Publicly embarrassed someone at work

1: Never	2: Seldom	3: Sometimes	4: Often	5: Always

Exit, voice, loyalty (EVLN; Farrell, 1983)

In regards to your current job, to what extent have you thought about engaging in the following behaviors in the last year:

- Exit
    - Getting into action and looking for another job
    - Deciding to quit the company
    - Getting myself transferred to another job

- Loyalty
    - Waiting patiently and hoping the problem will solve itself
    - Quietly doing my job and letting higher-ups make the decisions
    - Saying nothing to others and assuming things will work out
- Voice (Rusbult et al., 1988)
    - Going to my immediate supervisor to discuss the problem
    - Asking my co-workers for advice about what to do
    - Talking to the office manager about how I felt about the situation
    - Trying to solve the problem by suggesting changes in the way work was supervised in the office

1: Never	2: Seldom	3: Sometimes	4: Often	5: Always

Justice Measures (Colquitt, 2001)

Procedural justice (Colquitt, 2001; based on Thibaut & Walker, 1975 and Leventhal, 1980)

The following items refer to the procedures used to arrive at outcomes you receive from your job (e.g., pay, promotions, etc.). To what extent:

1. Have you been able to express your views and feelings during those procedures?
2. Have you had influence over the outcomes arrived at by those procedures?
3. Have those procedures been applied consistently?
4. Have those procedures been free of bias?
5. Have those procedures been based on accurate information?
6. Have you been able to appeal the outcomes arrived at by those procedures?
7. Have those procedures upheld ethical and moral standards?

Distributive justice (Colquitt, 2001; based on Leventhal, 1980)

The following items refer to outcomes you receive from your job (e.g., pay, promotions, etc.). To what extent:

1. Do your outcomes reflect the effort you have put into your work?
2. Are your outcomes appropriate for the work you have completed?
3. Do your outcomes reflect what you have contributed to the organization?
4. Are your outcomes justified, given your performance?

Interpersonal justice (Colquitt, 2001; based on Bies & Moag, 1986)

The following items refer to your supervisor. To what extent:

1. Have they treated you in a polite manner?
2. Have they treated you with dignity?
3. Have they treated you with respect?
4. Have they refrained from improper remarks or comments?

1: To a very small extent	2: To a small extent	3: To a moderate extent	4: to a large extent	5: To a very large extent

Social desirability (Hays et al., 1989)

Listed below are a few statements about your relationships with others. To what extent do you agree with each statement?

1. I am always courteous even to people who are disagreeable.
2. There have been occasions when I took advantage of someone.
3. I sometimes try to get even rather than forgive and forget.
4. I sometimes feel resentful when I don't get my way.
5. No matter who I'm talking to, I'm always a good listener.

1: Strongly Disagree	2: Disagree	3: Undecided	4: Agree	5: Strongly Agree

Neuroticism (Donnellan et al., 2006)

1. I have frequent mood swings.
2. I am relaxed most of the time. (R)
3. I get upset easily.
4. I seldom feel blue. (R)

1: Strongly Disagree	2: Disagree	3: Undecided	4: Agree	5: Strongly Agree

Problem-Focused Coping (Carver et al., 1989)

Please indicate what you usually do and feel, when you experience stressful events.

1. I try to come up with a strategy about what to do.
2. I force myself to wait for the right time to do something.
3. I put aside other activities in order to concentrate on this.
4. I make a plan of action.

1: Strongly Disagree	2: Disagree	3: Undecided	4: Agree	5: Strongly Agree

Trait Self-Control (Tagney et al., 2004)

Using the scale provided, please indicate how much each of the following statements reflects how you typically are.

1. I am good at resisting temptation.
2. I have a hard time breaking bad habits. (R)
3. I am lazy. (R)
4. I say inappropriate things. (R)
5. I do certain things that are bad for me, if they are fun. (R)
6. I refuse things that are bad for me.
7. I wish I had more self-discipline. (R)
8. People would say that I have iron self- discipline.
9. Pleasure and fun sometimes keep me from getting work done. (R)
10. I have trouble concentrating. (R)
11. I am able to work effectively toward long-term goals.
12. Sometimes I can’t stop myself from doing something, even if I know it is wrong. (R)
13. I often act without thinking through all the alternatives. (R)

1: Strongly Disagree	2: Disagree	3: Undecided	4: Agree	5: Strongly Agree

Demographic and work-related questions

- How comfortable are you with the English language (including speaking and writing)?
    1. Not at all comfortable
    2. A little comfortable
    3. Somewhat comfortable
    4. Comfortable
    5. Very comfortable
- What is your age (in years)?
    - Open-ended
- What is your gender identity?
    - Male
    - Female
    - Other
    - Prefer not to answer
- What is your race (select all that apply)?
    - Asian
    - Black/African American
    - Hispanic or Latinx
    - White/Caucasian
    - Other
- What is your marital status?
    - Single
    - Married
    - Widowed
    - Divorced
    - Separated
- What is your highest level of education?
    - No formal education credential
    - High school diploma or equivalent
    - Some college, no degree
    - Postsecondary nondegree award
    - Associate’s degree
    - Bachelor’s degree
    - Master’s degree
    - Doctoral or professional degree
- In what U.S. state or territory do you live?
    - List of all U.S. states and territories
- What is your approximate total yearly household income?
    - Less than $20,0000
    - $20,000-29,999
    - $30,000-39,999
    - $40,000-49,999
    - $50,000-59,999
    - $60,000-69,999
    - $70,000 or more
- What is your job title?

  - Open-ended response
- What industry do you work in?
  1. Admin and support
  2. Arts, entertainment, or recreation
  3. Aviation
  4. Construction
  5. Education
  6. Finance or insurance
  7. Food services
  8. Forestry, fishing, hunting or agriculture
  9. Health care or social assistance
  10. Information
  11. Management of companies or enterprises
  12. Manufacturing
  13. Military
  14. Mining
  15. Other
  16. Professional, scientific or technical services
  17. Public safety or emergency response
  18. Real estate or rental and leasing
  19. Retail trade
  20. Tourism or hospitality
  21. Transportation or warehousing
  22. Utilities
  23. Waste management or remediation services
  24. Wholesale trade
- How long have you worked in your current organization? For example, if you had worked 2 years and 3 months in your current organization, you would enter “2” in the years field and “3” in the months field.
  1. Open-ended (has 2 fields, one for year and one for number of month)
- How many hours do you work in an average week?
  - Open-ended response
- On average, what percentage of the workweek do you spend working remotely (away from your organization’s primary local office)?
  - Percentage slide from 0-100%

APPENDIX B: COMPUTERIZED TEXT ANALYSIS R CODE

Load packages

```
library(tidyverse) # Data wrangling
library(tm) # (1) lower case conversion, (2) removing non-alphabetic characters, and white space
library(hunspell) # Correct spelling.
library(qdap) # Expanding contractions and abbreviations.
library(textstem) # Lemmatization.
library(text2vec) # Text transformation
library(stringr) # String manipulation
library(stringi) # String manipulation
library(caret) # Machine learning
```

Read in data and create a variable with all text responses merged into one.

```
data <- read.csv("Data/Wave 1/clean.wave1.06.16.2021.csv") %>%
    mutate(full_text = paste(SJT_O1, SJT_O2, SJT_i1, SJT_i2, sep = " ")) %>%  # Merge all text responses into one variable.
```

Text preprocessing

Spell-Check

Create a vector with misspelled words.

```
bad_words <- hunspell(data$full_text,
                     ignore = c('hr', 'pto')) %>% # Add to dictionary
  unlist()
```

Grab the first suggestion for each misspelled word.

```
suggestions <- sapply(hunspell_suggest(bad_words), "[[", 1)
```

Compare the incorrectly spelled words with their suggested edits. Export this and manually double check how well the corrections of the misspelled words are and fix whatever needs fixing (e.g., in excel). Export this and import one that has been looked through.

```
spell_check <- cbind(bad_words, suggestions)
write.csv(spell_check, "spell_check.csv")
```

Read in the corrected words

```
spell_check <- read.csv("spell_check_fixed.csv")
```

Add spaces to bad words and suggestions. This is needed for doing regex later.

```
bad_words <- paste(" ", spell_check$bad_words, " ", sep = "")
suggestions <- paste(" ", spell_check$suggestions2, " ", sep = "") # suggestions2 is just a column of all the correct spellings, including the ones that were manually added.
```

Actually apply the spell-check and correct the spelling on the full_text, as well as on each individual CWB SJT item. While doing this, create a new column where the corrected text is inputted instead of overwriting the old text.

```
data <- mutate(data, clean_text = stringi::stri_replace_all_regex(
     str = full_text, pattern = bad_words, replacement = suggestions,
         stringi::stri_opts_fixed(case_insensitive = FALSE)))
```

Double check if there are any misspelled words left.

```
hunspell(data$clean_text, ignore = c()) %>%
  unlist()
```

All other preprocessing steps

Create a function that will be used to run the rest of the text pre-processing steps. These steps include (1) lower-case conversion, (2) expanding contractions, (3) adding work-related abbreviations, (4) lemmatization, (5) handle negation, (6) stop-word removal, and (7) removal of numbers, symbols, punctuation, and extra white space.

```
text_preprocessing <- function(string){

     # Lower-case conversion
     temp <- tolower(string)

     # Turn all "’" to "'" so that contractions are interpreted correctly.
     temp <- str_replace_all(temp, "’", "'")
     # Expand all contractions.
     temp <- replace_contraction(text.var = temp,
                      contraction = qdapDictionaries::contractions,
                      ignore.case = TRUE)

     # Create some common work-related abbreviations and make some
     # terminology more consistent to reduce the vocabulary size.
     abv <- c(" human resources ", " human resource ", " payed time off ",
          " co worker ", " co-worker "," ot ", " boss ")
     rep <- c(" HR ", " HR ", " PTO ", " coworker ", " coworker ", " OT ",
           " supervisor ")
     abbreviations <- as.data.frame(cbind(abv, rep))
     # Add abbreviations
     temp <- qdap::replace_abbreviation(text.var = temp,
          abbreviation = abbreviations, ignore.case = FALSE)

     # Lemmatization
     temp <- lemmatize_strings(temp)

     # Based on the negation words in `qdapDictionaries::negation.words`,
     # create a vector with the negation words and one with the the
     # negation words followed by '_'. Then apply the handle negation.
     negation <- c("not ", "never ", "no ", "nobody ", "nor ", "neither ")
```

```
    negation_fixed <- c("not_", "never_", "no_", "nobody_", "nor_",
            "neither_")
    temp <- stringi::stri_replace_all_regex(str = temp,
          pattern = negation, replacement = negation_fixed,
        stringi::stri_opts_fixed(case_insensitive = FALSE))

    # Make sure everything is lowercase after previous steps.
    temp <- tolower(temp)

    # Remove stop words
    temp <- removeWords(temp, words = stopwords::stopwords("en"))

    # Remove all numbers
    temp <- removeNumbers(temp, ucp = FALSE)

    # Remove punctuation
    temp <- removePunctuation(temp, ucp = FALSE,
                 preserve_intra_word_contractions = FALSE,
                 preserve_intra_word_dashes = TRUE)

    # Remove white space
    temp <- stripWhitespace(temp)
}
```

Process text responses using the text_preprocessing() function that was just created.

```
data$clean_text <- text_preprocessing(data$clean_text)
```

Preparing Outcome Variable

Create composite mean scores for CWB-I and CWB-O.

```
data$CWB_O <- rowMeans(data[,c('org_deviance_1', 'org_deviance_2',
     'org_deviance_3', 'org_deviance_4', 'org_deviance_5',
     'org_deviance_6', 'org_deviance_7', 'org_deviance_8',
     'org_deviance_9', 'org_deviance_10', 'org_deviance_11',
     'org_deviance_12')], na.rm = TRUE)

data$CWB_I <- rowMeans(data[,c('person_deviance_1', 'person_deviance_2',
        'person_deviance_3', 'person_deviance_4', 'person_deviance_5',
        'person_deviance_6', 'person_deviance_7')], na.rm = TRUE)
```

Create data frame with Mturk_ID and CWB_I and CWB_O (i.e., the labels).

```
CWB_O_label <- as.data.frame(cbind(data$Mturk_ID, data$CWB_O))
CWB_I_label <- as.data.frame(cbind(data$Mturk_ID, data$CWB_I))
```

Coerce MTurk_ID into a factor and CWB numerical scores from the Workplace Deviance Scale (Bennett & Robinson, 2000) to a numeric.

```
CWB_O_label[,1] <- as.factor(CWB_O_label[,1])
CWB_O_label[,2] <- as.numeric(CWB_O_label[,2])

CWB_I_label[,1] <- as.factor(CWB_I_label[,1])
CWB_I_label[,2] <- as.numeric(CWB_I_label[,2])
```

Rename the columns (V1 to Mturk_ID and V2 to label_T1)

```
CWB_O_label <- CWB_O_label %>% rename(Mturk_ID = V1, label_T1 = V2)
CWB_I_label <- CWB_I_label %>% rename(Mturk_ID = V1, label_T1 = V2)
```

Data Splitting

Create a 90/10 train/test data split.

```
# Set a random seed for reproducability.
set.seed(222)

# Split data into 90/10 split.
trainIndex <- createDataPartition(y = data$CWB_O, # Outcome variable.
                     p = .9, # Training percentage.
                     list = FALSE, # Output as matrix.
                     times = 1) # Number of partitions.
```

Split the data into train and test.

```
data_train <- data[ trainIndex,]
data_test  <- data[-trainIndex,]
```

Data transformation

Tokenize data.

```
tokens <- itoken(iterable = data$clean_text,
                 ids = data$Mturk_ID)

tokens_train <- itoken(iterable = data_train$clean_text,
                 ids = data_train$Mturk_ID)

tokens_test <- itoken(iterable = data_test$clean_text,
                 ids = data_test$Mturk_ID)
```

Create vocabulary based on unigram and bigrams.

```
vocab_2g_train  <- create_vocabulary(it = tokens_train,
             ngram = c(ngram_min = 1L, ngram_max = 2L),
             sep_ngram = "_")
```

Filter the vocabulary by removing words that
(1) occur less than 2 times through all documents,
(2) occur in maximum 99% of the documents, and
(3) occur in less than 2% of the documents.

```
pruned_vocab_2g_train = prune_vocabulary(vocab_2g_train,
                        term_count_min = 2,
                        doc_proportion_max = 0.99,
                        doc_proportion_min = 0.02)
```

Vectorize the pruned vocabulary.

```
vectorizer <- vocab_vectorizer(pruned_vocab_2g_train)
```

Create DTM in dgCMatrix form.

```
dtm_train <- create_dtm(it = tokens_train, vectorizer = vectorizer,
              type = "dgCMatrix")

dtm_test <- create_dtm(it = tokens_test, vectorizer = vectorizer,
                        type = "dgCMatrix")
```

Convert from sparse matrix to normal matrix and label columns.

```
matrix_train <- as.matrix(dtm_train)
colnames(matrix_train) <- colnames(dtm_train)

matrix_test <- as.matrix(dtm_test)
colnames(matrix_test) <- colnames(dtm_test)
```

Machine Learning

Create The Model

Set up 5-fold cross validation with 3 repeats.

```
cv5 <- trainControl(method = 'repeatedcv', number = 5, repeats = 3,
         savePredictions = TRUE)
```

Train LASSO model for CWB-O.

```
set.seed(222)

LASSO_fit_CWB_O <- train(x = matrix_train, y = CWB_O_label[trainIndex,2],
              method = 'lasso', trControl = cv5,
              tuneGrid = expand.grid(fraction = .7),
              metric = "RMSE"); LASSO_fit_CWB_O
```

Train LASSO model for CWB-I.

```
set.seed(222)

LASSO_fit_CWB_I <- train(x = matrix_train, y = CWB_I_label[trainIndex,2],
            method = 'lasso', trControl = cv5,
            tuneGrid = expand.grid(fraction = 1),
            metric = "RMSE"); LASSO_fit_CWB_I
```

Run Prediction

Create a function to predict on the test data and turn it into a clear format with Mturk_ID (i.e., ID variable) as a column.

```
data_prep_t1 <- function(string){
    pred <- predict(string, newdata = matrix_test)
    pred <- data.frame(pred)
    pred <- tibble::rownames_to_column(pred, var = 'Mturk_ID')
}
```

Use the created function on the LASSO models (i.e., CWB-O and CWB-I).

```
pred_LASSO_CWB_O <- data_prep_t1(LASSO_fit_CWB_O) %>%
  rename(pred_LASSO = pred)
pred_LASSO_CWB_I <- data_prep_t1(LASSO_fit_CWB_I) %>%
  rename(pred_LASSO = pred)
```

Create a data set with both the labels and predicted values of CWB.

```
pred_CWB_O <- dplyr::full_join(x = pred_LASSO_CWB_O, CWB_O_label,
          by = 'Mturk_ID', keep = F)
pred_CWB_I <- dplyr::full_join(x = pred_LASSO_CWB_I, CWB_I_label,
          by = 'Mturk_ID', keep = F)
```

Turn all values below 1 to 1.

```
pred_CWB_O$pred_LASSO <- ifelse(pred_CWB_O$pred_LASSO < 1 , 1, pred_CWB_O$pred_LASSO
)
pred_CWB_I$pred_LASSO <- ifelse(pred_CWB_I$pred_LASSO < 1 , 1, pred_CWB_I$pred_LASSO)
```

Evaluate Model

Check correlation between predicted values and labels of CWB, and examine RMSE, MAE, and R-squared.

```
# CWB-O
cor.test(pred_CWB_O$pred_LASSO, pred_CWB_O$label_T1)
postResample(pred = pred_CWB_O$pred_LASSO, obs = pred_CWB_O$label_T1)

```

```
# CWB-I
cor.test(pred_CWB_I$pred_LASSO, pred_CWB_O$label_T1)
postResample(pred = pred_CWB_I$pred_LASSO, obs = pred_CWB_I$label_T1)
```

APPENDIX C: STRUCTURAL TOPIC MODELING R CODE

Here it will be demonstrated how you can conduct structural topic modeling on the CWB SJT. This will include doing topic modeling on all the CWB SJT combined, as well as individually.

Load needed packages.

```
library(tidyverse) # Data management
library(hunspell) # Correct spelling
library(tm) # (1) Lower case conversion, (2) removing non-alphabetic characters, and white space
library(qdap) # Expanding contractions and abbreviations
library(textstem) # Lemmatization
library(text2vec) # Used for text transformation
library(stringr) # Manipulating strings
library(stringi) # Manipulating strings
```

Read in the data set. While doing this, create a variable with all text responses merged into one. Select and keep relevant variables.

```
data <- read.csv("CWB.SJT.dataset.csv") %>%
    mutate(full_text = paste(SJT_O1, SJT_O2, SJT_i1, SJT_i2, sep = " ")) %>%  # Merge all text respon
ses into one variable.
    select(Mturk_ID, full_text, SJT_O1, SJT_O2, SJT_i1, SJT_i2)
```

Text Preprocessing

Spell-Check

Create a vector with misspelled words.

```
bad_words <- hunspell(data$full_text,
                      ignore = c('hr', 'pto')) %>% # Words to add to dictionary
  unlist()
```

Grab the first suggestion for each misspelled word.

```
suggestions <- sapply(hunspell_suggest(bad_words), "[[", 1)
```

Compare the incorrectly spelled words with their suggested edits. Export this and manually double check how well the corrections of the misspelled words are and fix whatever needs fixing (e.g., in excel). Export this and import one that has been looked through.

```
spell_check <- cbind(bad_words, suggestions)
write.csv(spell_check, "spell_check.csv")
```

Read in the corrected words

```
spell_check <- read.csv("spell_check_fixed.csv")
```

Add spaces to bad words and suggestions. This is needed for doing regex later.

```
bad_words <- paste(" ", spell_check$bad_words, " ", sep = "")
suggestions <- paste(" ", spell_check$suggestions2, " ", sep = "") # suggestions2 is just a column of all th
e correct spellings, including the ones that were manually added.
```

Actually apply the spell-check and correct the spelling on the full_text, as well as on each individual CWB SJT item. While doing this, create a new column where the corrected text is inputted instead of overwriting the old text.

```
data <- data %>%
       mutate(clean_full_text = stri_replace_all_regex(
        str = full_text, pattern = bad_words, replacement = suggestions,
        stri_opts_fixed(case_insensitive = FALSE)),
             clean_sjt1 = stri_replace_all_regex(str = SJT_O1,
                   pattern = bad_words, replacement = suggestions,
                   stri_opts_fixed(case_insensitive = FALSE)),
             clean_sjt2 = stri_replace_all_regex(str = SJT_O2,
                        pattern = bad_words, replacement = suggestions,
                        stri_opts_fixed(case_insensitive = FALSE)),
             clean_sjt3 = stri_replace_all_regex(str = SJT_i1,
                        pattern = bad_words, replacement = suggestions,
                        stri_opts_fixed(case_insensitive = FALSE)),
             clean_sjt4 = stri_replace_all_regex(str = SJT_i2,
                        pattern = bad_words, replacement = suggestions,
                        stri_opts_fixed(case_insensitive = FALSE)))
```

Double check if there are any misspelled words left.

```
hunspell(data$clean_text, ignore = c()) %>%
  unlist()
```

All other preprocessing steps

Create a function that will be used to run the rest of the text pre-processing steps. These steps include (1) lower-case conversion, (2) expanding contractions, (3) adding work-related abbreviations, (4) lemmatization, (5) stop-word removal, and (6) removal of numbers, symbols, punctuation, and extra white space.

```
text_preprocessing <- function(string){

       # Lower-case conversion
       temp <- tolower(string)

       # Turn all "’" to "'" so that contractions are interpreted correctly.
       temp <- str_replace_all(temp, "’", "'")
       # Expand all contractions.
       temp <- replace_contraction(text.var = temp,
                     contraction = qdapDictionaries::contractions,
                     ignore.case = TRUE)


       # Create some common work-related abbreviations and make some
```

```
    # terminology more consistent to reduce the vocabulary size.
    abv <- c(" human resources ", " human resource ", " payed time off ",
     " co worker ", " co-worker "," ot ", " boss ")
    rep <- c(" HR ", " HR ", " PTO ", " coworker ", " coworker ", " OT ",
     " supervisor ")
    abbreviations <- as.data.frame(cbind(abv, rep))
    # Add abbreviations
    temp <- qdap::replace_abbreviation(text.var = temp,
         abbreviation = abbreviations, ignore.case = FALSE)

    # Lemmatization
    temp <- lemmatize_strings(temp)

    # Make sure everything is lowercase after previous steps.
    temp <- tolower(temp)

    # Remove stop words
    temp <- removeWords(temp, words = stopwords::stopwords("en"))

    # Remove all numbers
    temp <- removeNumbers(temp, ucp = FALSE)

    # Remove punctuation
    temp <- removePunctuation(temp, ucp = FALSE,
                  preserve_intra_word_contractions = FALSE,
                  preserve_intra_word_dashes = TRUE)

    # Remove white space
    temp <- stripWhitespace(temp)
}
```

Process text responses using the text_preprocessing() function that was just created.

```
data$clean_text <- text_preprocessing(data$clean_text)
data$clean_sjt1 <- text_preprocessing(data$clean_sjt1)
data$clean_sjt2 <- text_preprocessing(data$clean_sjt2)
data$clean_sjt3 <- text_preprocessing(data$clean_sjt3)
data$clean_sjt4 <- text_preprocessing(data$clean_sjt4)
```

Create separate data sets for each item and one for all items combined. Then remove NAs.

```
data_full_text <- data %>%
 select(Mturk_ID, full_text, clean_full_text, wave) %>% na.omit()
data_sjt1 <- data %>%
 select(Mturk_ID, SJT_O1, clean_sjt1, wave) %>% na.omit()
data_sjt2 <- data %>%
 select(Mturk_ID, SJT_O2, clean_sjt2, wave) %>% na.omit()
data_sjt3 <- data %>%
 select(Mturk_ID, SJT_i1, clean_sjt3, wave) %>% na.omit()
```

```
data_sjt4 <- data %>%
  select(Mturk_ID, SJT_i2, clean_sjt4, wave) %>% na.omit()
```

Take a look at the text before and after text preprocessing.

```
cbind(data$full_text, data$clean_text) %>%
   View()
```

Data Transformation

Create a function to create a document-term matrix based on text data.

```
data_transformation <- function(string, dataset) {

      # Tokenize text
      temp_token <- itoken(iterable = string,
                  ids = dataset$Mturk_ID)

      # Create a vocabulary using unigrams and bigrams
      vocab_temp  <- create_vocabulary(it = temp_token,
                        ngram = c(ngram_min = 1L, ngram_max = 2L),
                        sep_ngram = "_")

      # Prune vocabulary. Remove terms occurring less than 2 times, occur in        # more than 99% of doc
uments, and occur in less than 1% of documents.
      pruned_vocab_temp <- prune_vocabulary(vocab_temp,
                           term_count_min = 2,
                           doc_proportion_max = 0.98,
                           doc_proportion_min = 0.01)

      # Create DTM in dgCMatrix form.
      dtm <- create_dtm(it = temp_token,
                vectorizer = vocab_vectorizer(pruned_vocab_temp),
                type = "dgCMatrix")

      # Remove rows in the DTM where all values are 0 due to pruning.
      sel_idx <- slam::row_sums(dtm) > 0
      dtm <- dtm[sel_idx, ]
}
```

Create another similar function that outputs data set with respondents who have their entire DTM row being 0 after pruning removed.

```
data_transformation_data_output <- function(string, dataset) {

      # Tokenize text
      temp_token <- itoken(iterable = string,
                              ids = dataset$Mturk_ID)
```

```
    # Create a vocabulary using unigrams and bigrams
    vocab_temp  <- create_vocabulary(it = temp_token,
                ngram = c(ngram_min = 1L, ngram_max = 2L),
                sep_ngram = "_")

    # Prune vocabulary. Remove terms occurring less than 2 times, occur in     # more than 99% of documents, and occur in less than 1% of documents.
    pruned_vocab_temp <- prune_vocabulary(vocab_temp,
                  term_count_min = 2,
                  doc_proportion_max = 0.98,
                  doc_proportion_min = 0.01)

    # Create DTM in dgCMatrix form.
    dtm <- create_dtm(it = temp_token,
           vectorizer = vocab_vectorizer(pruned_vocab_temp),
           type = "dgCMatrix")

    # Remove rows in the DTM where all values are 0 due to pruning.
    sel_idx <- slam::row_sums(dtm) > 0
    dtm <- dtm[sel_idx, ]
    dataset <- dataset[sel_idx, ]
}
```

Create a document-term matrix for each item and the combined items.

```
dtm_full_text <- data_transformation(data_full_text$clean_full_text, dataset = data_full_text)
dtm_sjt1 <- data_transformation(data_sjt1$clean_sjt1, dataset = data_sjt1)
dtm_sjt2 <- data_transformation(data_sjt2$clean_sjt2, dataset = data_sjt2)
dtm_sjt3 <- data_transformation(data_sjt3$clean_sjt3, dataset = data_sjt3)
dtm_sjt4 <- data_transformation(data_sjt4$clean_sjt4, dataset = data_sjt4)
```

Create separate data sets and remove rows where DTM values are 0.

```
data_full_text <- data_transformation_data_output(data_full_text$clean_full_text, dataset = data_full_text)
data_sjt1 <- data_transformation_data_output(data_sjt1$clean_sjt1, dataset = data_sjt1)
data_sjt2 <- data_transformation_data_output(data_sjt2$clean_sjt2, dataset = data_sjt2)
data_sjt3 <- data_transformation_data_output(data_sjt3$clean_sjt3, dataset = data_sjt3)
data_sjt4 <- data_transformation_data_output(data_sjt4$clean_sjt4, dataset = data_sjt4)
```

Structural Topic Modeling

Full Data

First, the full text (i.e., including responses on all 4 SJT items) will be examined.

Create corpus from the DTM of the full text.

```
corp <- readCorpus(dtm_full_text, type = "dtm")
```

Next, specify meta variables.

```
data_full_text$wave <- as.factor(data_full_text$wave)
meta_vars <- data_full_text[,c('wave', 'Mturk_ID')] # Just adding Mturk_ID because by only inputting 'wave' later functions don't work as 'wave' will lose its variable name. Any variable can be added in place of Mturk_ID here.
```

Prep documents

```
out <- prepDocuments(documents = corp$documents, vocab = corp$vocab,
                     meta = meta_vars, lower.thresh = 2)
```

Find the ideal number of topics (K).

```
# Set seed for reproducibility of results
set.seed(222)

# Find ideal topic number for the data
stm_search <- searchK(documents = out$documents, vocab = out$vocab,
                  K = 2:20, # Topics 2 through 20 will be assessed.
                  init.type = "Spectral",
                  prevalence = ~ wave, data = out$meta, verbose = FALSE)
```

Check for Semantic Coherence and Exclusivity to see what topic number is best to use.

```
ggplot(data = as.data.frame(stm_search$results),
     aes(x = as.numeric(semcoh), y = as.numeric(exclus))) +
  geom_text(aes(label = K), show.legend = F, check_overlap = F, size = 3.6,
      family = "Times New Roman") +
  labs(x = 'Semantic coherence', y = 'Exclusivity') +
  scale_x_continuous(limits = c(-60, -30),
         breaks = c(-60, -55, -50, -45, -40, -35, -30)) +
  scale_y_continuous(limits = c(7.3, 9.2)) +
  theme_classic() +
  theme(text = element_text(size = 12,  family = "Times New Roman"))
```

Plot more metrics for diagnosing the most appropriate number of topics.

```
plot(stm_search)
```

Run model with chosen number of topics.

```
stm <- stm(documents = out$documents, vocab = out$vocab,
                  K = 5, # The number of topics
                  init.type = "Spectral",
                  prevalence = ~ wave, data = out$meta,
                  seed = 222, verbose = FALSE)
```

Get top 10 words for each topic.

```
labelTopics(stm, n = 10)
```

Visualize topics

```
plot(stm, type = "labels", main = "Topic terms")
plot(stm, n = 7, text.cex = .8)
```

Examine example responses that fit under each topic.

```
findThoughts(stm, texts = data_full_text$full_text,
             n = 1, topics = c(1:5))
```

CWB SJT Item 1

Next, the same process is applied to do topic modeling on the individual SJT items.

Create a corpus.

```
corp_sjt1 <- readCorpus(dtm_sjt1, type = "dtm")
```

Specify meta variables.

```
data_sjt1$wave <- as.factor(data_sjt1$wave)
meta_vars_sjt1 <- data_sjt1[,c('wave', 'Mturk_ID')]
```

Prep documents.

```
out_sjt1 <- prepDocuments(documents = corp_sjt1$documents,
                vocab = corp_sjt1$vocab,
                meta = meta_vars_sjt1, lower.thresh = 2)
```

Find the ideal topic number (K).

```
set.seed(222)
stm_search_sjt1 <- searchK(documents = out_sjt1$documents,
           vocab = out_sjt1$vocab,
                   K = 2:20, # Assess topics 2 through 20.
                   init.type = "Spectral",
                   prevalence = ~ wave, data = out_sjt1$meta,
           verbose = FALSE)
```

Check for Semantic Coherence and Exclusivity to see what topic number is best to use.

```
ggplot(data = as.data.frame(stm_search_sjt1$results),
      aes(x = as.numeric(semcoh), y = as.numeric(exclus))) +
   geom_text(aes(label = K), show.legend = F, check_overlap = F, size = 3.6,
      family = "Times New Roman") +
   labs(x = 'Semantic coherence', y = 'Exclusivity') +
   scale_x_continuous(limits = c(-60, -30),
          breaks = c(-60, -55, -50, -45, -40, -35, -30)) +
   scale_y_continuous(limits = c(7.3, 9.2)) +
   theme_classic() +
   theme(text = element_text(size = 12,  family = "Times New Roman"))
```

Plot more metrics for diagnosing the most appropriate number of topics.

```
plot(stm_search_sjt1)
```

Run model with chosen number of topics.

```
stm_sjt1 <- stm(documents = out_sjt1$documents, vocab = out_sjt1$vocab,
           K = 4, # The number of topics
               init.type = "Spectral",
               prevalence = ~ wave, data = out_sjt1$meta,
               seed = 222, verbose = FALSE)
```

Get top 10 words for each topic.

```
labelTopics(stm_sjt1, n = 10)
```

Visualize topics

```
plot(stm_sjt1, type = "labels", main = "Topic terms")
plot(stm_sjt1, n = 7, text.cex = .8)
```

Examine example responses that fit under each topic.

```
findThoughts(stm_sjt1, texts = data_sjt1$SJT_O1,
          n = 1, topics = c(1:4))
```

CWB SJT Item 2-4

To do topic modeling for items 2-4, the same exact process that was used to do topic modeling on the first item can be used.

APPENDIX D: SUPPLEMENTAL TOPIC MODELING RESULTS

Topic modeling results for each individual CWB SJT item will be discussed here.

SJT Item 1

Models with topic numbers two through twenty were examined to find the ideal number of topics to model for the first SJT item. Four topics was indicated as the most appropriate number of topics in the full data set containing wave one and two data, as it maximized semantic coherence and exclusivity scores, as well as parsimony (see Figure 4). As an additional measure of examining the ideal number of topics, individual metrics for semantic coherence, held-out likelihood, lower bound, and residuals for each model were also examined (see Figure 5). The finding from this also suggested that four topics to be ideal. For the top ten words represented under each of these topics—as judged by a FREX score—please see Table 4. The first topic included words relating to guessing and submitting work, the second included terms relating to coworkers and asking for help, the third topic had words relating to guessing, and the fourth topic was more ambiguous but had terms that were present in the situational judgment item itself.

The change in topics from wave one to wave two was also examined. Findings indicate that the expected topic proportions for three of the four topics were represented to a highly comparable extent in both waves (see Figure 6). The other topic (topic 2) did seem to be more represented in the first wave. Overall, as three of the four topics were represented to a similar extent in both waves of data collection, it provides some reliability evidence.

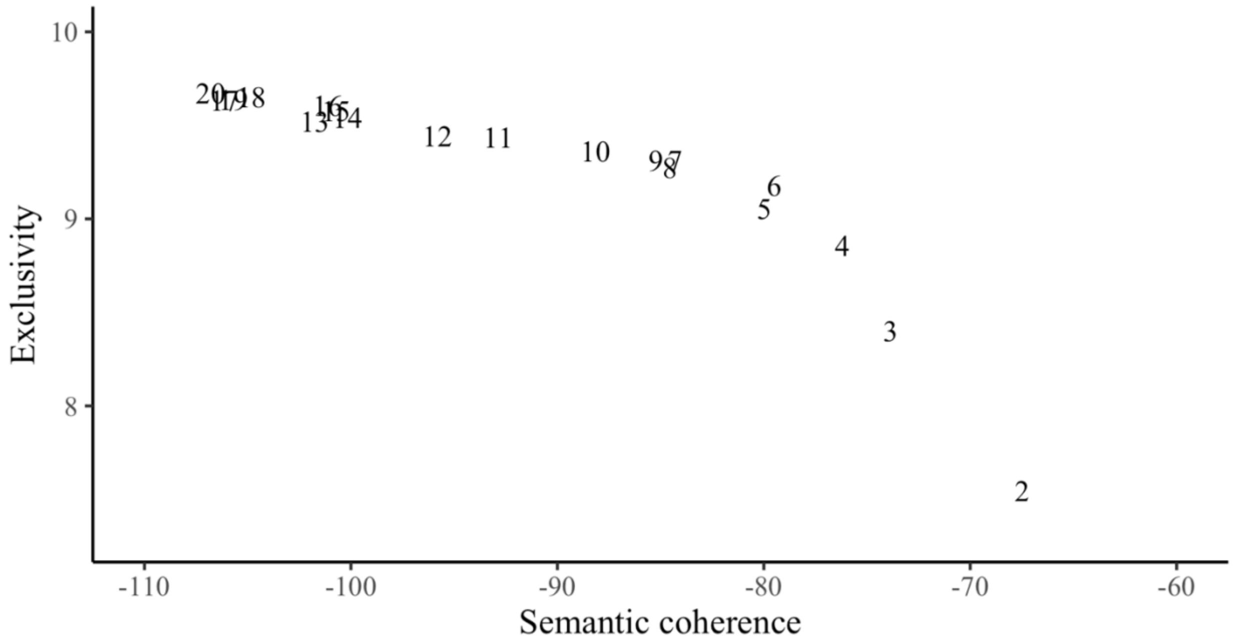


Figure 4
Semantic Coherence and Exclusivity Scores for Topics 2 Through 20 for SJT Item 1

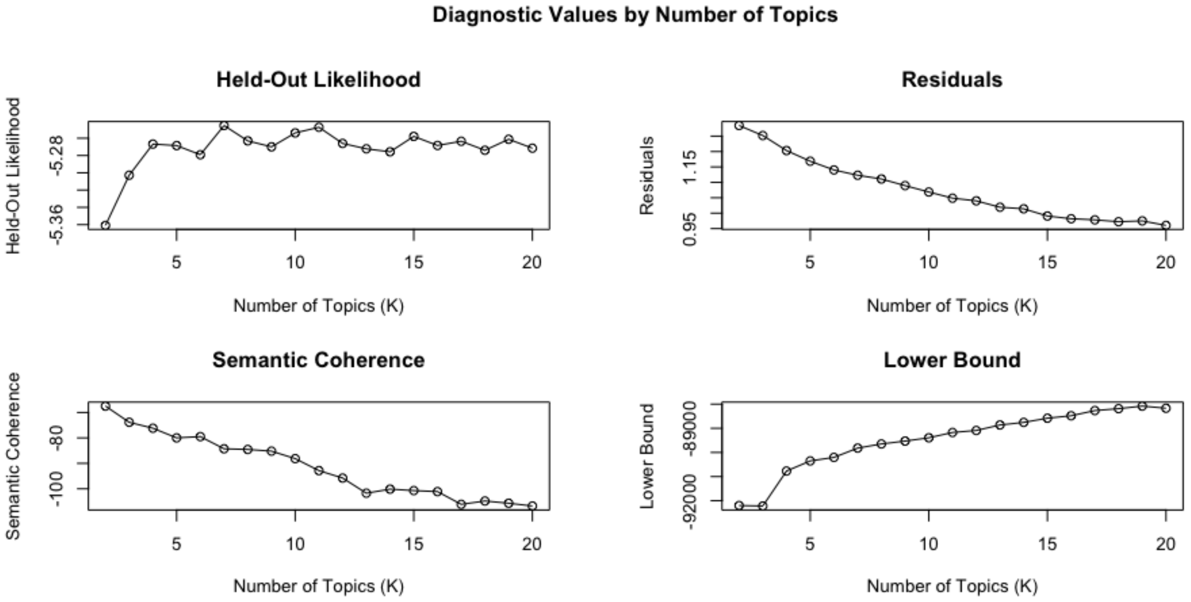


Figure 5
Evaluation Metrics of Topic Models with Topics Ranging from 2 to 20 for SJT Item 1

Table 4

Topics and Top Words Representing Them According to FREX Scores for SJT Item 1

Topics			
Topic 1	Topic 2	Topic 3	Topic 4
guess_rest	help	make_sure	estimate
turn_something	ask_help	good_can	datum
performance_bonus	coworker	make	information
rest	_ask	_make	provide
performance	stay	good_guess	earnings
something	ask	sure	quarterly
_guess	team	make_good	quarterly_earnings
turn	supervisor	good	client
rest_calculation	task	_good	available
risk	coworker_help	educate_guess	report

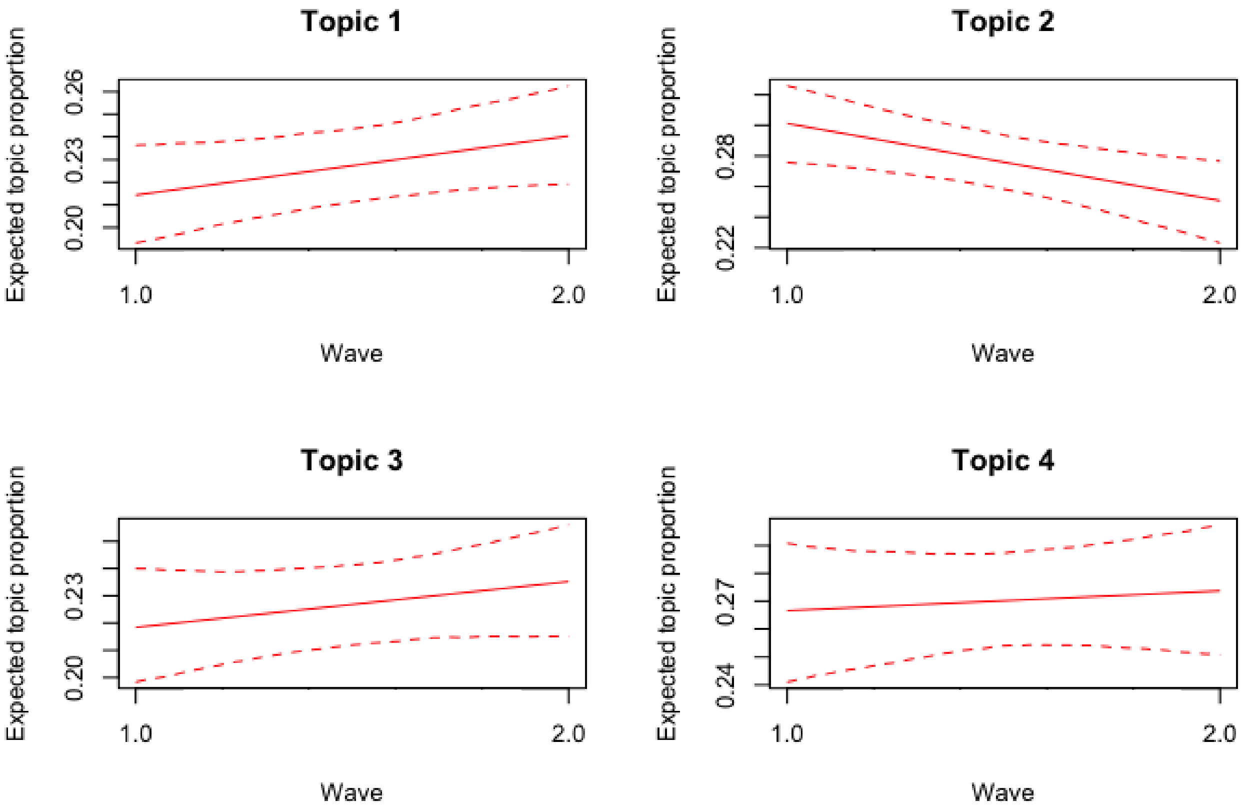


Figure 6

Change in Topic Prevalence from Wave 1 to Wave 2 for Each Topic for SJT Item 1

SJT Item 2

Models with topic numbers two through twenty were examined to find the ideal number of topics to model for the second SJT item. The most appropriate number of topics was not immediately apparent from FREX scores, as it suggested that 3, 4, 5, or 7 topics might be ideal (see Figure 7). As a next step, semantic coherence, held-out likelihood, lower bound, and residuals for each model were also examined. These metrics indicated 5 topics as being most appropriate (See Figure 8). However, upon examining individual topics for each model, the author found that four topic model made more sense due to clearer and more distinguished topics; as such, this model was used. For the top ten words represented under each of these four topics—as judged by a FREX score—please see Table 5. The first topic involved words relating to communication, the second included terms relating to workload and communication, the third topic had words relating to performance, and the fourth topic included terms related to looking for a new job.

The change in topics from wave one to wave two was also examined. Findings indicated that the expected topic proportions for each topic were represented to a highly comparable extent in both waves (see Figure 9). This provides reliability evidence as the same topics are mentioned when the item is measured at different time points.

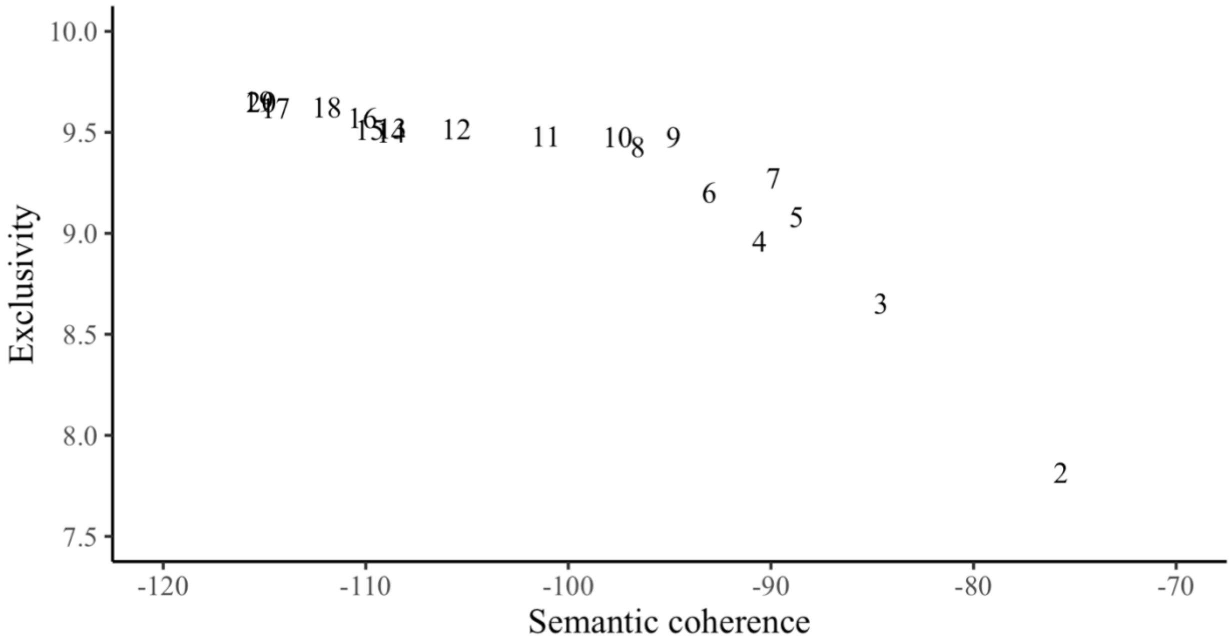


Figure 7
Semantic Coherence and Exclusivity Scores for Topics 2 Through 20 for SJT Item 2

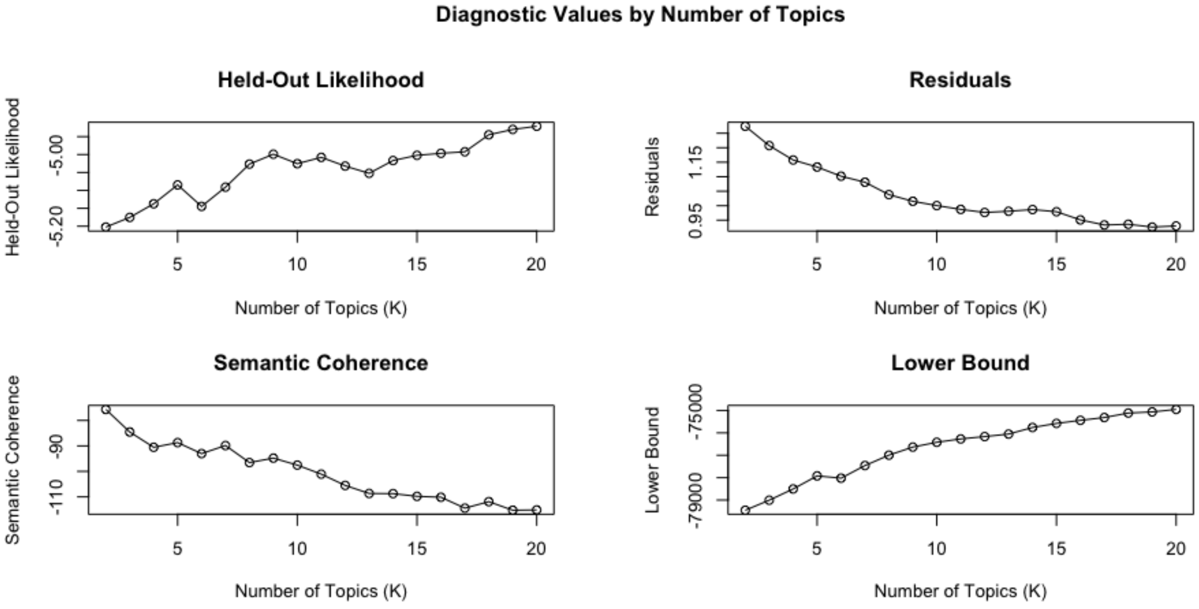


Figure 8
Evaluation Metrics of Topic Models with Topics Ranging from 2 to 20 for SJT Item 2

Table 5

Topics and Top Words Representing Them According to FREX Scores for SJT Item 2

Topics			
Topic 1	Topic 2	Topic 3	Topic 4
supervisor	much_work	hard	_ask
feel	_go	_continue	another_job
talk	go_supervisor	work_hard	look_another
talk_supervisor	extra_work	slow	start_look
_talk	much_pay	continue	meet_supervisor
appreciate	supervisor_explain	keep	new_job
can	receive	hard_work	meet
tell	_speak	continue_work	look_new
know	pay_much	day	ask_raise
situation	much	time	ask_supervisor

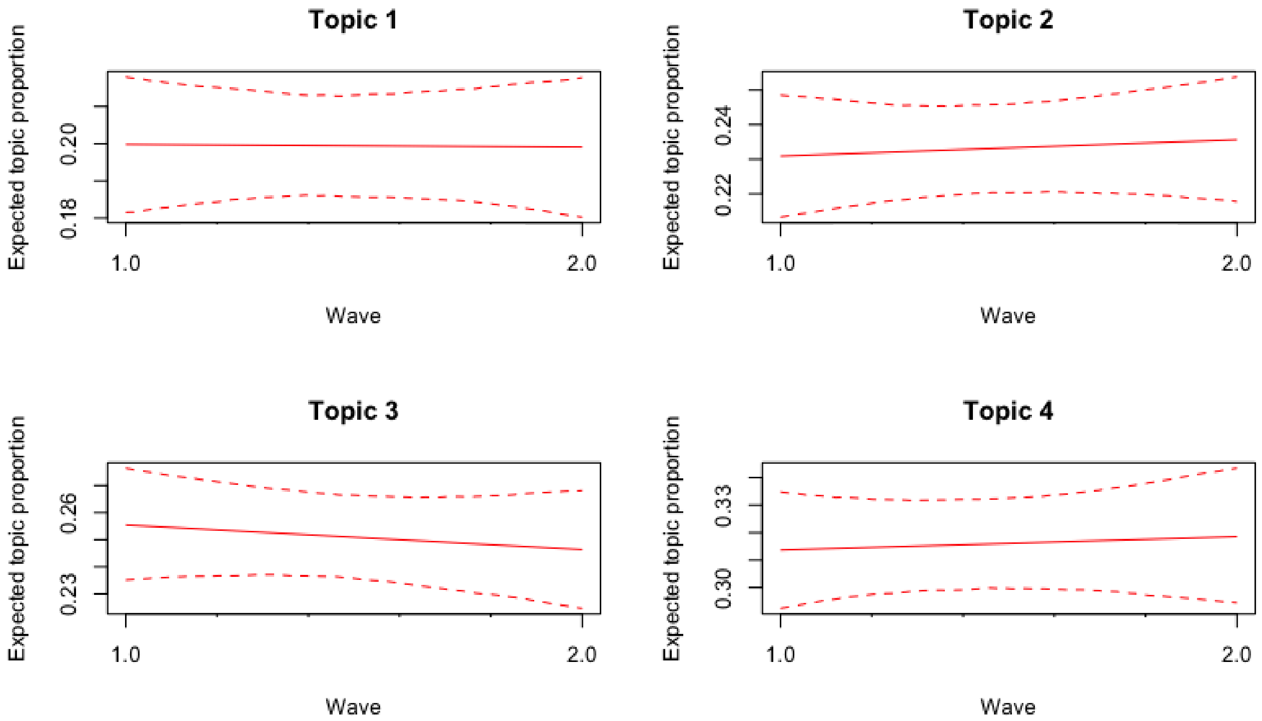


Figure 9

Change in Topic Prevalence from Wave 1 to Wave 2 for Each Topic for SJT Item 2

SJT Item 3

Models with topic numbers two through twenty were examined to find the ideal number of topics to model for the third SJT item. Four topics was indicated as the most appropriate number of topics in the full data set containing wave one and two data, as it maximized semantic coherence and exclusivity scores, as well as parsimony (see Figure 10). As an additional way of examining the ideal topic number, semantic coherence, held-out likelihood, lower bound, and residuals for each model were examined. Findings from this indicated that five topics might be more appropriate (see Figure 11). Due to the conflicting suggestions, the author examined the topics for each model and concluded that four topics was more appropriate as the topics were clearer. For the top ten words represented under each of these topics—as judged by a FREX score—please see Table 6. The first topic is fairly ambiguous and does not seem to represent a clear, specific topic. In contrast, the other topics are more distinguished: the second topic includes words relating to communication, the third topic comprises terms relating to project completion and complaints, and the fourth topic includes words related to confrontation.

The change in topics from wave one to wave two responses was also examined. Findings indicated that the topics were largely represented to the same extent in both waves (see Figure 12). This provides reliability evidence for the third SJT item.

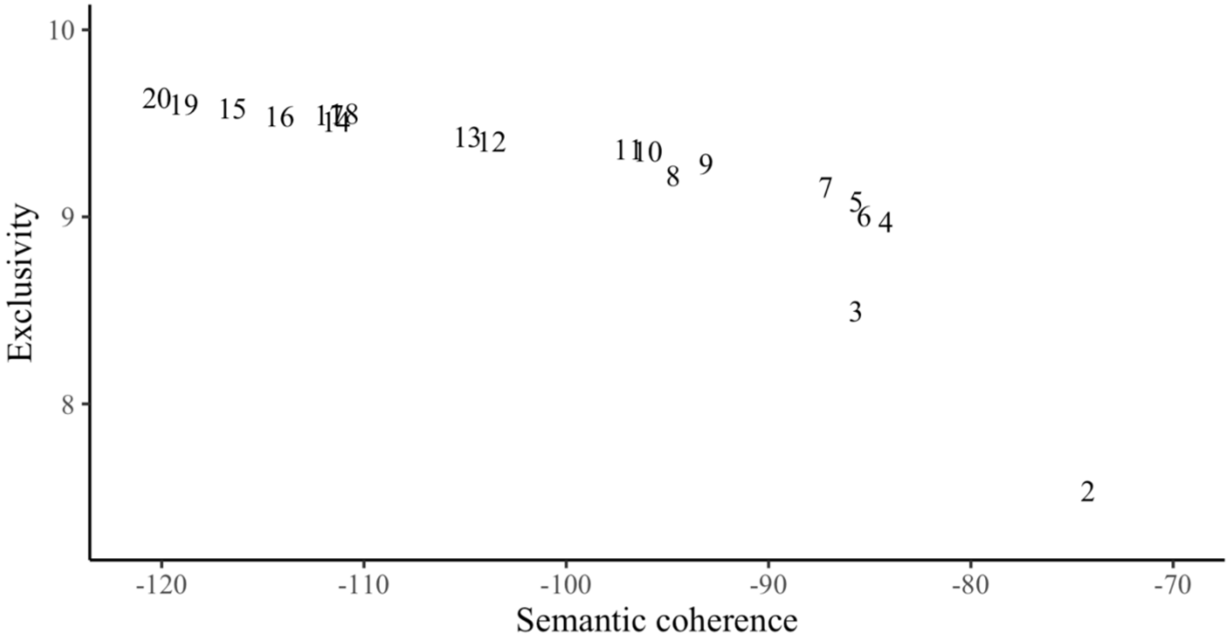


Figure 10
Semantic Coherence and Exclusivity Scores for Topics 2 Through 20 for SJT Item 3

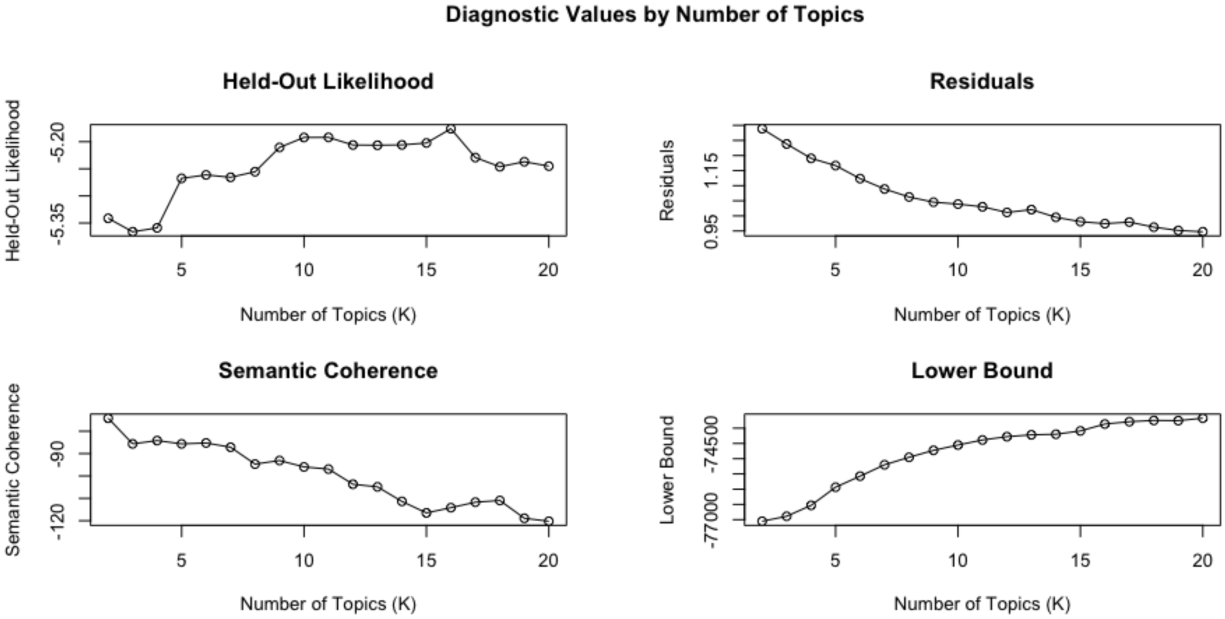


Figure 11
Evaluation Metrics of Topic Models with Topics Ranging from 2 to 20 for SJT Item 3

Table 6

Topics and Top Words Representing Them According to FREX Scores for SJT Item 3

Topics			
Topic 1	Topic 2	Topic 3	Topic 4
day	tell_coworker	complaint	confront
environment	_tell	finish	_confront
next_day	let_know	finish_project	problem
treat	know	complete	just
report_hr	talk_coworker	stand	_ask
take	tell	yell_back	ask
next	explain	_finish	rude
work_environment	supervisor_explain	file_complaint	think
hostile	walk	file	confront_coworker
continue_work	_go	coworkers	want

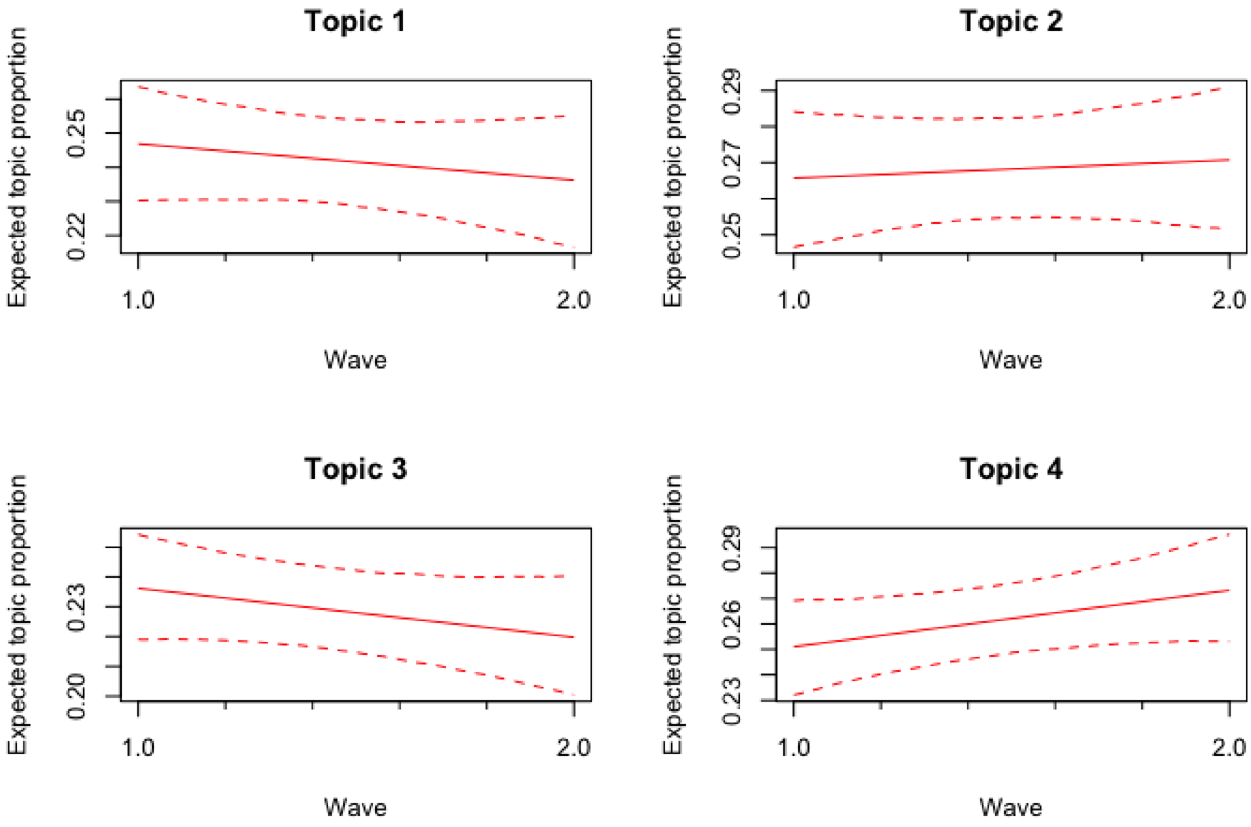


Figure 12

Change in Topic Prevalence from Wave 1 to Wave 2 for Each Topic for SJT Item 3

SJT Item 4

Models with topic numbers two through twenty were examined to find the ideal number of topics to model for the fourth and last SJT item. Semantic coherence, exclusivity, and parsimony were maximized for models with five, four, and three topics (see Figure 13). As an additional way of examining the ideal topic number, semantic coherence, held-out likelihood, lower bound, and residuals for each model were examined. Findings from this indicate that five topics might be most appropriate (see Figure 14). As an extra step, the author examined the models with three, four, and five topics and concluded that three topics was most appropriate as the topics were clearer than for the other models. For the top ten words represented under each of these topics—as judged by a FREX score—please see Table 7. The first topic included terms related to looking for a new job, the second topic had words about money and earnings, and the third topic had terms related to communication.

The change in topics from wave one to wave two responses was also examined. Findings indicated that the topics were largely represented to the same extent in both waves (see Figure 15). This provides reliability evidence for the third SJT item.

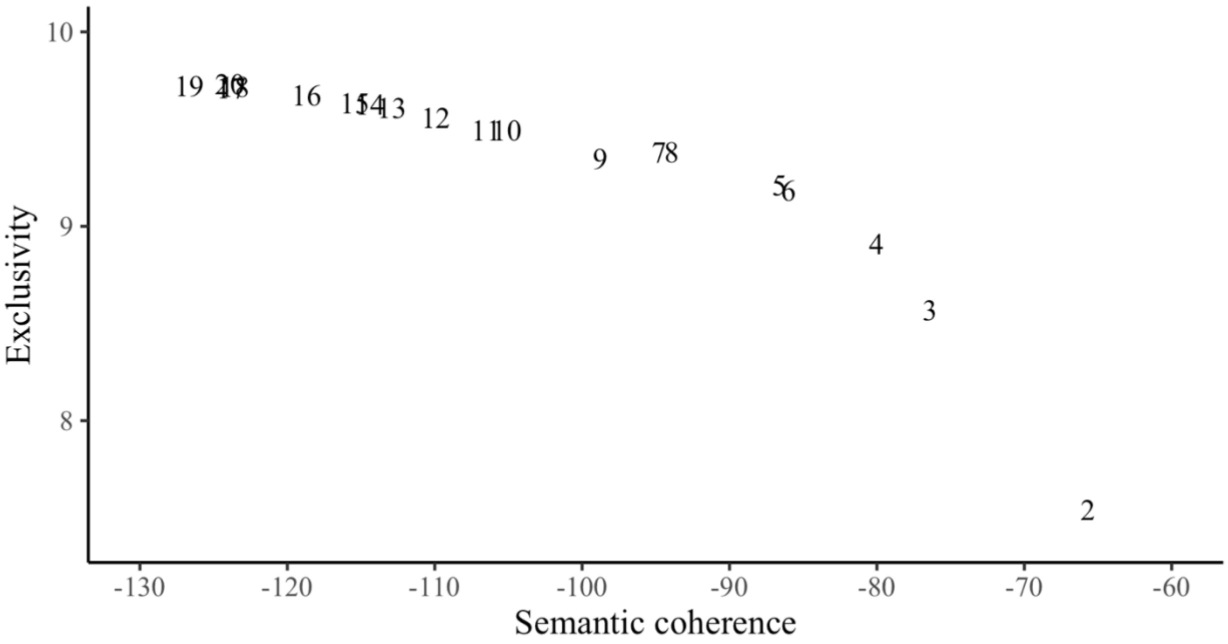


Figure 13
Semantic Coherence and Exclusivity Scores for Topics 2 Through 20 for SJT item 4

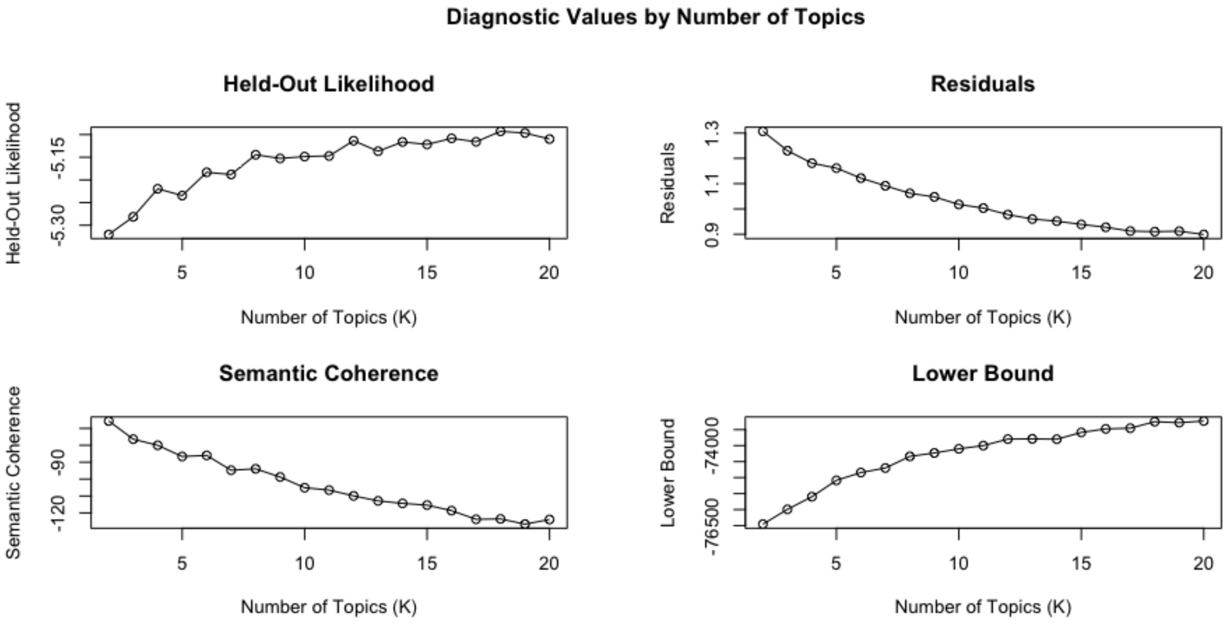


Figure 14
Evaluation Metrics of Topic Models with Topics Ranging from 2 to 20 for SJT Item 4

Table 7

Topics and Top Words Representing Them According to FREX Scores for SJT Item 4

	Topics	
Topic 1	Topic 2	Topic 3
start	money	talk_supervisor
_handle	much_money	_ask
start_look	make_much	get_raise
another_job	much	talk
look_another	earn_much	tell_supervisor
ability	employee_make	get_pay
look	earn	go
good_ability	much_work	approach
new_job	little_experience	tell
project good	make	bring

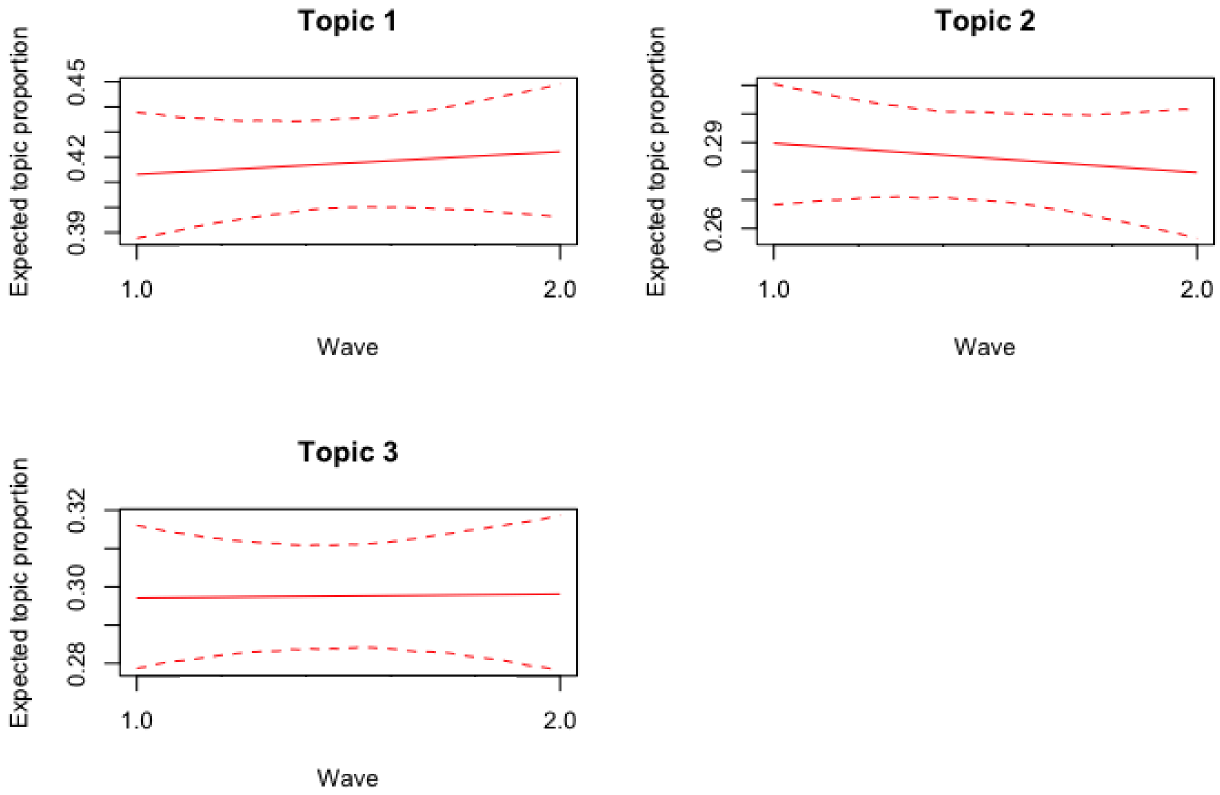
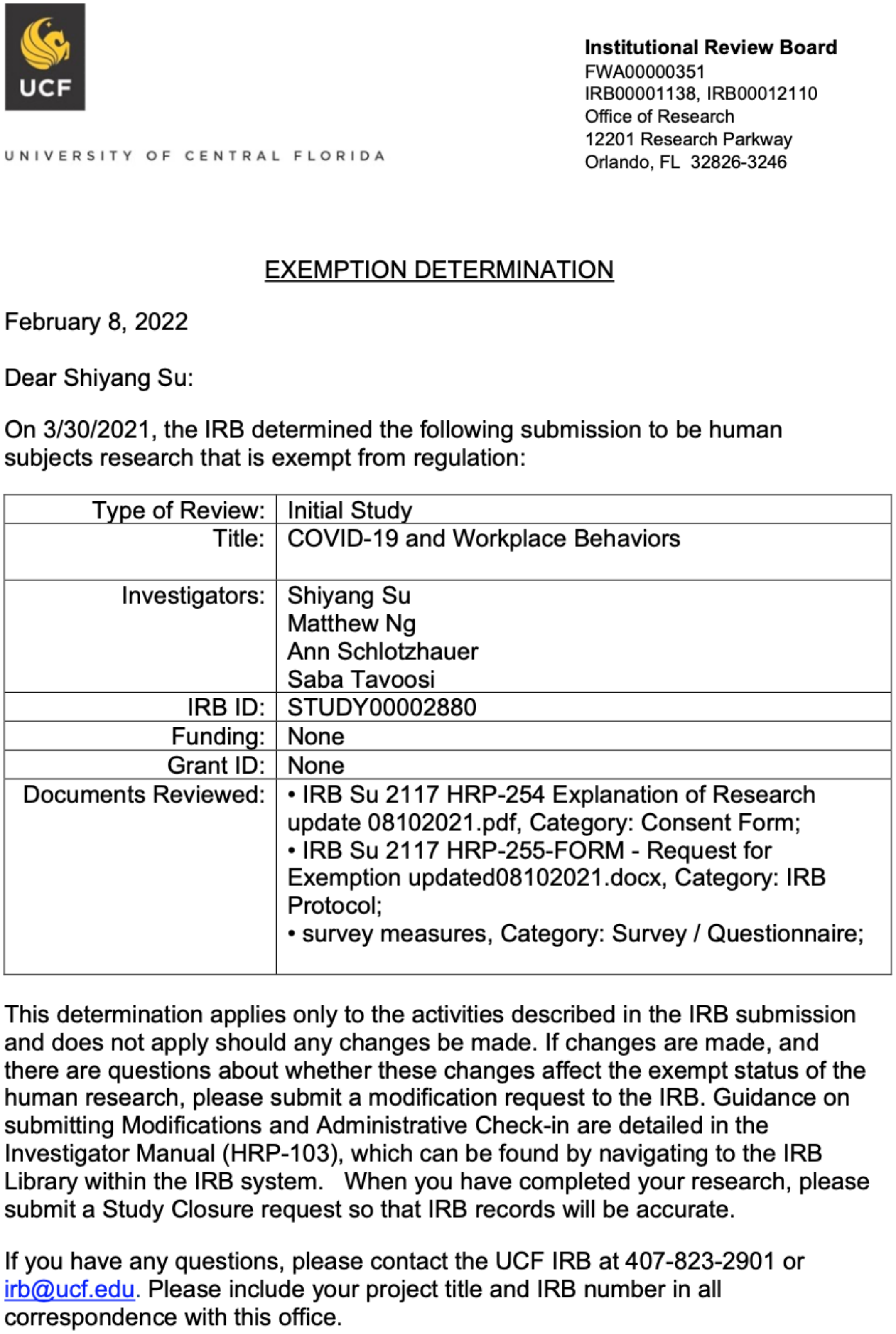


Figure 15

Change in Topic Prevalence from Wave 1 to Wave 2 for Each Topic for SJT Item 4

APPENDIX E: IRB APPROVAL LETTER



UCF

UNIVERSITY OF CENTRAL FLORIDA

Institutional Review Board
FWA00000351
IRB00001138, IRB00012110
Office of Research
12201 Research Parkway
Orlando, FL 32826-3246

EXEMPTION DETERMINATION

February 8, 2022

Dear Shiyang Su:

On 3/30/2021, the IRB determined the following submission to be human subjects research that is exempt from regulation:

Type of Review:	Initial Study
Title:	COVID-19 and Workplace Behaviors
Investigators:	Shiyang Su Matthew Ng Ann Schlotzhauer Saba Tavoosi
IRB ID:	STUDY00002880
Funding:	None
Grant ID:	None
Documents Reviewed:	• IRB Su 2117 HRP-254 Explanation of Research update 08102021.pdf, Category: Consent Form; • IRB Su 2117 HRP-255-FORM - Request for Exemption updated08102021.docx, Category: IRB Protocol; • survey measures, Category: Survey / Questionnaire;

This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made, and there are questions about whether these changes affect the exempt status of the human research, please submit a modification request to the IRB. Guidance on submitting Modifications and Administrative Check-in are detailed in the Investigator Manual (HRP-103), which can be found by navigating to the IRB Library within the IRB system.   When you have completed your research, please submit a Study Closure request so that IRB records will be accurate.

If you have any questions, please contact the UCF IRB at 407-823-2901 or irb@ucf.edu. Please include your project title and IRB number in all correspondence with this office.

Sincerely,

Renea Carver

Renea Carver
UCF IRB

REFERENCES

Adams, J. S. (1965). Inequity in social exchange. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2). New York: Academic Press.

Arun, r., Suresh , V., Veni Madhavan, C. E., Narasimha Murthy, M. N. (2010). On finding the natural number of topics with latent Dirichlet allocation: Some observations. In Advances in knowledge discovery and data mining, Zaki, M. J., Yu, J. X., Ravindran, B., and Pudi, V. (eds.). Springer Berlin Heidelberg, 391–402.

Becker, T. E. (2005). Development and validation of a situational judgment test of employee integrity. *International Journal of Selection & Assessment, 13*(3), 225–232.

Bennett, R. J., & Robinson, S. L. (2000). Development of a measure of workplace deviance. *Journal of Applied Psychology, 85*(3), 349–360.

Bennett, R. J., Marasi, S., & Locklear, L. (2019). Workplace deviance. In *Oxford research encyclopedia of business and management: 1-25*. New York: Oxford University Press.

Berry, C. M., Carpenter, N. C., & Barratt, C. L. (2012). Do other-reports of counterproductive work behavior provide an incremental contribution over self-reports? A meta-analytic comparison. *Journal of Applied Psychology, 97*(3), 613–636.

Berry, C. M., Ones, D. S., & Sackett, P. R. (2007). Interpersonal deviance, organizational deviance, and their common correlates: A review and meta-analysis. *Journal of Applied Psychology, 92,* 410–424.

Bies, R. J., & Moag, J. *S.* (1986). Interactional justice: Communication criteria of fairness. In R. J. Lewiclu, B. H. Sheppard, & M. H. Bazerman (Eds.), *Research on negotiation organizations* (Vol. 1, pp. 43-55). Greenwich, CT: JAI.

Blau, G. (1995). Influence of group lateness on individual lateness: A cross level examination. *Academy of Management Journal, 38*, 1483−1495.

Bledow, R., & Frese, M. (2009). A situational judgment test of personal initiative and its relationship to performance. *Personnel Psychology, 62*(2), 229–258.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *Annals of Applied Statistics, 1*(1), 17-35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(1), 993-1022.

de Boer, B. J., van Hooft, E. A. J., & Bakker, A. B. (2015). Self-control at work: Its relationship with contextual performance. *Journal of Managerial Psychology, 30*(4), 406-421.

Bolton, L. R., Becker, L. K., & Barber, L. K. (2010). Big Five trait predictors of differential counterproductive work behavior dimensions. *Personality and Individual Differences, 49*(5), 537–541.

Bowling, N. A., & Beehr, T. A. (2006). Workplace harassment from the victim's perspective: A theoretical model and meta-analysis. *Journal of Applied Psychology, 91*(5), 998–1012.

Bowling, N. A., & Gruys, M. L. (2010). *Overlooked issues in the conceptualization and measurement of counterproductive work behavior. Human Resource Management Review, 20(1), 54–61.*

Budd, J. W., Arvey, R. D., & Lawless, P. (1996). Correlates and conse- quences of workplace violence. *Journal of Occupational Health Psy- chology, 1*, 197–210.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, *6*(1), 3–5.

Burns, G. N., & Christiansen, N. D. (2006). Sensitive or senseless: On the use of social desirability measures in selection and assessment. In R. L. Griffith & M. H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 115–150). Greenwich, CT: Information Age.

Camara, W. J., & Schneider, D. L. (1994). Integrity tests: Facts and unresolved issues. *American Psychologist*, *49*, 112−119.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81-105.

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.

Campion, M. C., Ployhart, R. E., & MacKenzie, W. I., Jr. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance, 27*(4), 283–310.

Carver, C. S., Scheier, M. F., & Weintraub, J. K. (1989). Assessing coping strategies: a theoretically based approach. *Journal of personality and social psychology, 56*(2), 267–283.

Chan, D. & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82*(1), 143-159.

Chen, E. E., & Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological Methods, 21*(4), 458-474.

Chen, P. Y. & Spector, P. E. (1992). Relationships of work stressors with aggression, withdrawal, theft and substance use: An exploratory study. *Journal of Occupational and Organizational Psychology, 65,* 177-184.

Ciarlante, K. (2019) *Conceptualizing the role of severity in counterproductive work behavior: Predicting employee engagement in minor and severe CWBs* (Publication No. 6468). [Master's thesis, University of Central Florida]. STARS Electronic Theses and Dissertations, 2004-2019.

Colquitt, J. A. (2001). On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology, 86*(3), 386–400.

Cullen, M. J., Sackett, P. R., & Lievens, F. (2006). Threats to the Operational Use of Situational Judgment Tests in the College Admission Process. *International Journal of Selection and Assessment, 14*(2), 142–155.

Dahlke, J. A., & Sackett, P. R. (2017). The relationship between cognitive ability saturation and subgroup mean differences across predictors of job performance. *Journal of Applied Psychology, 102,* 1403–1420.

Dalal, R. S. (2005). A Meta-Analysis of the Relationship Between Organizational Citizenship Behavior and Counterproductive Work Behavior. *Journal of Applied Psychology, 90*(6), 1241–1255.

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

DeMore, S. W, Fisher, J. D., & Baron, R. M. (1988). The equity-control model as a predictor of vandalism among college students. *Journal of Applied Social Psychology, 18,* 80-91.

Deutsch, M. (1985). *Distributive justice: A social psychological perspective.* New Haven, CT: Yale University Press.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP Scales: Tiny-yet-effective measures of the Big Five Factors of Personality. *Psychological Assessment, 18*(2), 192–203.

Downer, K., Wells, C., & Crichton, C. (2019). All work and no play: A text analysis. *International Journal of Market Research*, *61*(3), 236-251.

Dunlop, P. D., & Lee, K. (2004). Workplace deviance, organizational citizenship behavior, and business unit performance: the bad apples do spoil the whole barrel. *Journal of Organizational Behavior, 25*(1), 67–80.

Edward, B. D., & Arthur, Jr., W. (2007). An examination of factors contributing to a reduction in subgroup differences on a constructed-response paper-and-pencil test of scholastic achievement. *Journal of Applied Psychology, 92*(3), 794-801.

Efron, B., & Hastie, T. (2016). *Computer age statistical inference*. New York, NY: Cambridge University Press.

Ekehammar, B. (1974). Interactionism in personality from a historical perspective. *Psychological Bulletin, 81*(12), 1026–1048.

Esuli, A., & Sebastiani, F. (2010). Machines that learn how to code open-ended survey data. *International Journal of Market Research*, *52*(6), 775-800.

Farrell, D. (1983). Exit, voice, loyalty, and neglect as responses to job dissatisfaction: A multidimensional scaling study. *Academy of Management Journal, 26*(4), 596–607.

Feinerer, I. (2018). Introduction to the tm package: Text mining in R. Retrieved from https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software, 25*(5), 1–54.

Finch, H., Finch, M. E. H., Mcintosh, C. E., & Braun, C. (2018). The use of topic modeling with latent Dirichlet analysis with open-ended survey items. *Translational Issues in Psychological Science, 4*(4), 403-424.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1-22.

Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1-22.

Funke, U. & Schuler, H. (1998). Validity of stimulus and response components in a video test of social competence. *International Journal of Selection and Assessment, 6*(2), 115-123.

Gosling, S. D., Rentfrow, P. J., & Swann Jr., W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*, 504-528.

Greco, L. M., O'Boyle, E. H., & Walter, S. L. (2015). Absence of malice: A meta-analysis of nonresponse bias in counterproductive work behavior research. *Journal of Applied Psychology, 100*(1), 75–97.

Greenberg, J. (1990). Employee theft as a reaction to underpayment inequity: The hidden cost of pay cuts. *Journal of Applied Psychology, 75*(5), 561–568.

Greenberg, J., & Scott, K. S. (1996). Why do workers bite the hand that feeds them? Employee theft as a social exchange process. In B. M. Staw & L. L. Cummings (Eds.), *Research on organizational behavior* (Vol. 18, pp. 111-156). Greenwich, CT JAI Press.

Grubb III, W. L. (2003). *Situational judgment and emotional intelligence tests: Constructs and faking*. Unpublished doctoral dissertation. Virginia Commonwealth University, Richmond, VA.

Grubb, W. L., III (2005). *The Emotional Side of a Situational Judgement Test*. From online journal ‘Business Quest’. Available at westga.edu/~bquest/2005/emotional (November 8, 2019).

Grün, B., Hornik, K. (2011). Topicmodels: An R package for fitting topic models. *Journal of Statistical Software, 40*(13), 1-30.

Guenole, N., Chernyshenko, O. S., & Weekly, J. (2017). On Designing Construct Driven Situational Judgment Tests: Some Preliminary Recommendations. International Journal of Testing,17(3), 234-252.

Guo, F., Gallagher, C. M., Sun, T., Tavoosi, S., & Min, H. (2021). Smarter people analytics with organizational text data: Demonstrations using classic and advanced NLP models. *Human Resource Management Journal*. Advance Online Publication.

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence, 1*(1), 60-76.

Guzzo, R., & Carlisle, T. (2014). *Big data: Catch the wave*. Workshop presented at annual conference of the Society for Industrial and Organizational Psychology, Honolulu, HI.

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. Psychological Methods, 21(4), 447-457.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York: NY: Springer.

Hays, R. D., Hayashi, T., & Stewart, A. L. (1989). A five-item measure of socially desirable response set. *Educational and Psychological Measurement, 49*(3), 629–636.

Heggestad, E. D., Scheaf, D. J., Banks, G. C., Hausfeld, M. M., Tonidandel, S., & Williams, E. B. (2019). Scale adaption in organizational science research: A review and best-practice recommendations. *Journal of Management, 35*(6), 2596-2627.

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2020). Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations. *Organizational Research Methods.*

Hirschman, A. O. (1970). *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states.* Cambridge, MA: Harvard University Press.

Hoerl, A., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

Hollinger, R. C., & Clark, J. P. (1983). *Theft by employees.* Lexington, MA: Lexington Books.

Hooper, A. C., Cullen, M. J., & Sackett, P. R. (2006). *Operational Threats to the Use of SJTs: Faking, Coaching, and Retesting Issues.* In J. A. Weekley & R. E. Ployhart (Eds.), *SIOP organizational series. Situational judgment tests: Theory, measurement, and application* (p. 205–232). Lawrence Erlbaum Associates Publishers.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effects of response distortion on those validities. *Journal of Applied Psychology Monograph, 75,* 581–595.

Huang, H. & Zhang, B. (2009). Text segmentation. In L. Liu & M. T. Ozsu (Eds.), Encyclopedia of database systems (pp. 3072-3075). New York, NY: Springer. Retrieved from http://link.springer.com/referenceworkentry/10.1007/978-0-387-39940-9_421

IBM. (2012). *IBM SPSS Modeler Text Analytics 15 user's guide*. Armonk, NY: Author.

Iliev, R., Deghani, M., Sagi, E. (2015). Automated text analysis in psychology: methods, applications, and future developments. *Language and Cognition, 7*, 265-290.

In'nami, Y. & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing, 26*(2), 219-244.

Jermier, J. M., Knights, D., & Nord, W. R. (Eds.). (1994). *Critical perspectives on work and organization. Resistance and power in organizations*. Taylor & Frances/Routledge.

Jeroen Ooms (2018). hunspell: High-performance stemmer, tokenizer, and spell checker. R package version 3.0. https://CRAN.R-project.org/package=hunspell

Kanning, U. P., Grewe, K., Hollenberg, S., & Hadouch, M. (2006). From the subjects' point of view. *European Journal of Psychological Assessment, 22*, 168–176.

Kao, A. & Poteet, S. R. (2007). *Natural language processing and text mining*. London, United Kingdom: Springer, London.

Kjell, Kjell, Garcia, & Sikström. (2019). Semantic Measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods, 4*(1), 92-115.

Knorz, C., & Zapf, D. (1996). Mobbing—eine extreme Form sozialer Stressoren am Arbeitsplatz [Mobbing: A severe form of social stressors at work]. *Zeitschrift für Arbeitsund Organisationspsychologie, 40*(1), 12–21.

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

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

Kozako, I. N. ‘A. M. F., Safin, S. Z., & Rahim, A. R. A. (2013). The relationship of big five personality traits on counterproductive work behavior among hotel employees: An exploratory study. *Procedia Economics and Finance, 7,* 181-187.

Lenhard, W., Baier, H., Hoffmann, J., & Schneider, W. (2007). Automatische bewertung offener antworten mittels latenter semantischer analyse [Automatic scoring of constructed-response items with latent semantic analysis]. *Diagnostica, 53*, 155–165.

Leventhal, G. S., Karuza, J., & Fry, W. R. (1980). Beyond fairness: A theory of allocation preferences. In G. Mikula (Ed.), *Justice and social interaction* (pp. 167-218). New York: Springer-Verlag.

Lievens, F. (2017). Construct-Driven SJTs: Toward an Agenda for Future Research. *International Journal of Testing*, 17(3), 269-276.

Lievens, F., & Sackett, P. R. (2007). Situational judgment tests in high-stakes settings: Issues and strategies with generating alternate forms. *Journal of Applied Psychology, 92*(4), 1043–1055.

Lievens, F., Buyse, T., Sackett, P. R., & Connelly, B. S. (2012). The effects of coaching on situational judgment tests in high-stakes selection. *International Journal of Selection and Assessment, 20*(3), 272–282.

Lievens, F., Sackett, P. R., Dahlke, J. A., Oostrom, J. K., & De Soete, B. (2019). Constructed response formats and their effects on minority-majority differences and validity. *Journal of Applied Psychology, 104*(5), 715-726.

Liu, M., & Wronski, L. (2017). Examining completion rats in web surveys via over 25,000 real-world surveys. *Social Science Computer Review, 36*(1), 116-124.

Mangione, T. W., & Quinn, R. P. (1975). Job satisfaction, counterproductive behavior and drug use at work. *Journal of Applied Psychology, 60*, 114−116.

Marcus, B., Taylor, O. A., Hastings, S. E., Sturm, A., & Weigelt, O. (2016). The structure of counterproductive work behavior: A review, a structural meta-analysis, and a primary study. *Journal of Management, 42*(1), 203–233.

Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9*, 103−113.

McDaniel, M. A., & Whetzel, D. L. (2005). Situational judgment test research: Informing the debate on practical intelligence theory. *Intelligence, 33*(5), 515–525.

McDaniel, M. A., & Whetzel, D. L. (2007). *Situational judgment tests.* In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (p. 235–257). Taylor & Francis Group/Lawrence Erlbaum Associates.

McDaniel, M. A., & Whetzel, D. L., & Nguyen, N. T. (2006). *Situational judgment tests in personnel selection: A monograph for the International Personnel Management Association Assessment Council*. Alexandria, VA: International Personnel Management Assessment Council.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. III. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60*(1), 63–91.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*(4), 730-740.

Motowidlo, S. J. & Tippins, N. (1993). Further studies of the low-fidelity simulation in the form of a situational inventory. *Journal of Occupational and Organizational Psychology, 66*(4), 337-344.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*(6), 640–647.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006a). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of applied psychology*, *91*(4), 749–761.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006b). *A Theoretical Basis for Situational Judgment Tests.* In J. A. Weekley & R. E. Ployhart (Eds.), *SIOP organizational series. Situational judgment tests: Theory, measurement, and application* (p. 57–81). Lawrence Erlbaum Associates Publishers.

Mumford, T. V., Van Iddekinge, C. H., Morgeson, F. P., & Campion, M. A. (2008). The Team Role Test: Development and validation of a team role knowledge situational judgment test. *Journal of Applied Psychology, 93*(2), 250–267.

Mussel, P., Gatzka, T., & Hewig, J. (2018). Situational judgment tests as an alternative measure for personality assessment. *European Journal of Psychological Assessment, 34*(5), 328–335.

Nguyen, N. (2001). *Faking in situational judgment tests: An empirical investigation of the work problems survey.* Unpublished doctoral dissertation. Virginia Commonwealth University, Richmond, VA.

Nikita, M. (2019). ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters. https://CRAN.R-project.org/package=ldatuning

Nunnally, J. C., Jr. (1970). *Introduction to psychological measurement.* McGraw-Hill.

Olson-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. D., Keenan, P. A., & Donovan, M. A. (1998). Interactive video assessment of conflict resolution skills. *Personnel Psychology, 51*(1), 1–24.

Oostrom, J. K., de Vries, R. E., & de Wit, M. (2019). Development and validation of a HEXACO situational judgment test. *Human Performance, 32*(1), 1-29.

Pandey, S., & Pandey, S. K. (2019). Applying Natural Language Processing Capabilities in Computerized Textual Analysis to Measure Organizational Culture. *Organizational Research Methods*, *22*(3), 765–797.

Peterson, M. H., Griffith, R. L., Isaacson, A. J., O'Connell, M. S., & Mangos, P. M. (2011). Applicant faking, social desirability, and the prediction of counterproductive work behaviors. *Human Performance, 24*(3), 270-290.

Peus, C., Braun, S., & Frey, D. (2013). Situation-based measurement of the full range of leadership model—Development and validation of a situational judgment test. *The Leadership Quarterly, 24*(5), 777–795.

Pietsch, A.-S. & Lessman, S. (2018). Topic modeling for analyzing open-ended survey responses. *Journal of Business Analytics, 1*(2), 93-116.

Ployhart, R. E., & Ehrhart, M. G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection and Assessment, 11*(1), 1–16.

Poeppelman, T., Blacksmith, N., & Yang, Y. (2013). "Big data" technologies: Problem or solution? *The Industrial-Organizational psychologist, 51,* 119–126.

Putka, D. J., Beatty, A. S., & Reeder, M. C. (2018). Modern Prediction Methods: New Perspectives on a Common Problem. *Organizational Research Methods*, *21*(3), 689–732.

R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, *18,* 25–39.

Rinker, T. W. (2018). textstem: Tools for stemming and lemmatizing text version 0.1.4. Buffalo, New York. http://github.com/trinker/textstem

Rinker, T. W. (2020). qdap: Quantitative Discourse Analysis Package. 2.4.2. Buffalo, New York. https://github.com/trinker/qdap

Roberts, M. E., Brandon M. S., Dustin T., & Edoardo M. A. (2013). The structural topic model and applied social science. *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*.

Roberts., M. E, Stewards, B. M., & Tingley, D. (2019). STM: An R package for structural topic models. *Journal of Statistical Software, 91*(2), 1-40.

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

Robie, C., Komar, S., & Brown, D. J. (2010). The effects of coaching and speeding on Big Five and Impression Management Scale scores. *Human Performance, 23*(5), 446–467.

Robinson, S. L., & Bennett, R. J. (1995). A typology of deviant workplace behaviors: A multidimensional scaling study. *Academy of Management Journal, 38*(2), 555–572.

Rusbult, C., Farrell, D., Rogers, G., & Mainous, A. (1988). Impact of exchange variables on exit, voice, loyalty, and neglect: An integrative model of responses to declining job satisfaction. *Academy of Management, 31,* 599-627.

Sackett, P.R. and Devore, C.J. (2001) Counterproductive Behaviors at Work. In: Anderson, N., Ones, D., Sinangil, H. and Viswesvaran, C., Eds., Handbook of Industrial, Work, and Organizational Psychology, Sage, London, 145-164.

Salgado, J. F., Viswesvaran, C., & Ones, D. S. (2001). Predictors used for personnel selection: An overview of constructs, methods and techniques. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work and organizational psychology, Volume 1: Personnel psychology.* (pp. 165–199). Thousand Oaks, CA: Sage Publications Ltd.

Schmiedel, T., Müller, O., & vom Brocke, J. (2019). Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture. *Organizational Research Methods, 22*(4), 941–968.

Schmitt, N., & Chan, D. (2006). *Situational Judgment Tests: Method or Construct?* In J. A. Weekley & R. E. Ployhart (Eds.), *SIOP organizational series. Situational judgment tests:*

Theory, measurement, and application (p. 135–155). Lawrence Erlbaum Associates Publishers.

Selivanov, D., Bickel, M., & Wang, Q. (2020). Text2vec: Modern text mining framework for R. https://CRAN.R-project.org/package=text2vec

Sharma, S., Gangopadhyay, M., Austin, E., & Mandal, M. K. (2013). Development and validation of a situational judgment test of emotional intelligence. *International Journal of Selection and Assessment*, *21*(1), 57–73.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. *International Encyclopedia of Education*, *4*, 20–26.

Silge, J., Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in R. *Journal of Open Source Software, 1*(3), 37.

Skarlicki, D. P., & Folger, R. (1997). Retaliation in the workplace: The roles of distributive, procedural, and interactional justice. *Journal of Applied Psychology, 82*(3), 434–443.

Smith, D. B., & Ellingson, J. E. (2002). Substance vs. style: A new look at social desirability in motivating contexts. *Journal of Applied Psychology, 87*, 211–219.

Smith, K. C. & McDaniel, M. A. (1998). *Criterion and construct validity evidence for a situational judgment measure.* Paper presented at the 13th Annual Convention of the Society for Industrial and Organizational Psychology, Dallas, TX.

Snyder, M., & Ickes, W. (1985). Personality and social behavior. In G. Lindzey, & E. Aronson (Eds.), *Handbook of social psychology* (3rd Edition ed., pp. 883-948). Random House.

Spector, P. E., & Fox, S. (2005). The stressor–emotion model of counterproductive work behavior. In S. Fox & P. Spector (Eds.), *Counterproductive work behavior:*

Investigations of actors and targets (pp. 151–174). Washington, DC: American Psychological Association.

Spector, P. E., Fox, S., Penney, L. M., Bruursema, K., Goh, A., & Kessler, S. (2006). The dimensionality of counterproductivity: Are all counterproductive behaviors created equal? *Journal of Vocational Behavior, 68,* 446–460.

Stevens, M. J., & Campion, M. A. (1999). Staffing work teams: Development and validation of a selection test for teamwork settings. *Journal of Management, 25*(2), 207–228.

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality, 72*, 271-324.

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88*(3), 500-517.

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality, 34*(4), 397–423.

Thompson, I. (2019). *SIOP Machine Learning 2019 Data* [Jupyter Notebook]. Retrieved from https://github.com/izk8/2019_SIOP_Machine_Learning_Winners (Original work published 2018)

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, B, 58*(1), 267-288.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133,* 859–883.

Wagner, R. K., & Sternberg, R. J. (1991). Tacit Knowledge Inventory for Managers. San Antonio, TX: Psychological Corporation.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09). Association for Computing Machinery, New York, NY, USA, 1105–1112.

Weekley, J. A, & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52*(3), 679–700.

Weiss, H. M., & Adler, S. (1984). Personality and organizational behavior. *Research in Organizational Behavior, 6,* 1–50.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52*(5), 372–376.

Westring, A. J. F., Oswald, F. L., Schmitt, N., Drzakowski, S., Imus, A., Kim, B., & Shivpuri, S. (2009). Estimating trait and situational variance in a situational judgment test. *Human Performance, 22*(1), 44–63.

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review, 19*(3), 188–202.

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance, 21*(3), 291–309.

Withey, M. J., & Hooper, W. H. (1989). Predicting exit, voice, loyalty, and neglect. *Administrative Science Quarterly, 34*(4), 521-539.

Zapf, D., Knorz, C., & Kulla, M. (1996). On the relationship between mobbing factors, and job content, social work environment, and health outcomes. *European Journal of Work and Organizational Psychology, 5*(2), 215–237.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530.

Zhang, Y., Chen, M., & Liu, L. (2015). A review on text mining, In 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China: IEEE.