BI Norwegian Business School - campus Oslo

GRA 19502

Master Thesis

Component of continuous assessment: Thesis Master of Science
Final master thesis – Counts 80% of total grade

Leadership and Semantics: Explorations of the Semantic Theory of Survey Responses

Navn:	Emily Clare Noack, Cecilie Sell Bonde

Start:	02.03.2018 09.00
Finish:	03.09.2018 12.00

Abstract

In the field of organizational psychology, it has been found that responses to leadership surveys can consistently be predicted when a semantic analysis is run on the questions themselves. This indicates that leadership surveys are not contributing to our understanding of leadership or leadership attitudes, but are instead measuring something else. Using the Semantic Theory of Survey Responses, we first explore how semantic processing and emotionally-laden judgements occur in the brain separately in order to lay the foundation for a pilot fMRI study. Secondly, we tested STSR by survey manipulation to discover the resulting effects when respondents are forced to take more time to consider MLQ survey questions. We found that when individuals were forced to pause and think before answering, they become even more semantically driven and better at categorizing items into the five transformational leadership factors. These findings have implications for research on leadership, survey construction and analysis, and future fMRI studies.

Contents

Abstract 1
Introduction 3
Surveys 4
Semantic Theory of Survey Responses 5
Neuroscience of Semantics 7
Neuroscience of Emotions and Attitudes 9
Current fMRI Methods 10
Behavioral Methods of Testing STSR 11
Manipulation of Temporal Characteristics 12
Methods 13
Participants 13
Experiment 13
Analysis 15
Item-distances (190 item distance pairs) 15
Language parsing algorithms 15
Item-distance groups 16
Item distance analysis 17
Factor analysis 18
Correlation analysis 18
Results 19
Discussion 24
Limitations 27
Conclusion 28
References 30
Appendix A 36
Appendix B 37

Introduction

There are various theories arguing for one definition of leadership over another, and yet we find ourselves using circular explanations all the same (Eddy & VanDerLinden, 2006). In fact, the words "leader" and "leadership" have different, nuanced meanings depending on the different operating language (Schedlitzki, Ahonen, Wankhade, Edwards, & Gaggiotti, 2017). Though we use the word "leadership" in many settings--for example in the workplace, in schools, in sports, in politics--we rarely pause to contemplate if we truly know what exactly such a word entail.

How academic research on leadership translates to a practical setting can be even more complicated. Researchers Avolio and Hannah (2008) observe that there is not one accepted theory about how to develop leaders. Neither is there a clear understanding about the organizational or individual factors that facilitate or accelerate such development. Yet millions of dollars are invested in leadership development programs in corporate and educational entities (Riggio, 2008). For example, in a study completed by Lunsford and Brown (2017) which studied 69 collegiate leadership centers, annual budgets ranged from $1,500 to $900,000, with near two thirds of the centers having a budget of $100,000 or more. Further investigation led to the conclusion that the implementation of evidence-based practices were not reflected in the leadership centers' programs, philosophies or mission statements. One has to wonder if wasted effort and money on leadership and leadership development is due to our failure to solve the basic problem of what leadership is and what it means to people.

Majority of the research on leadership and leadership styles has been conducted through the use of field surveys (Yukl, 2013). For example, the Multifactor Leadership Questionnaire (Bass & Avolio, 1990) is considered the standard instrument for assessing a range of transformational, transactional, and non leadership scales (Rowold, 2005). However, as researchers Podsakoff, MacKenzie and colleagues (2003) point out, surveys are often part of the problem when it comes to common method bias, especially due to item embeddedness. In their critical literature review, they cite research by Harrison and McLaughlin (1993), which further describes item-embeddedness as when the respondent, after analyzing clues from context, use an easily accessible set of cognitions to answer subsequent items. Could these context clues be the logic, linguistics, and

meanings of the words used in the questions themselves? In line with the research of Arnulf and colleagues (2014), we think the answer could be 'yes'.

The purpose of this paper is to examine whether it is possible to design an experiment that allows us to study the difference between semantic and emotional determinants of survey responses. To accomplish this, we will first review the literature on leadership surveys and how the semantics of survey questions have potentially skewed our understanding of the concept itself. We will then discuss some of the empirical studies examining the cognitive processes that take place when subjects fill out such surveys, and whether it is possible to contribute to the theory using fMRI technique. Finally, we will conduct a behavioral experiment to test the STSR by specifically manipulating temporal aspects of surveys, which in theory should have no effect on semantic predictability of responses.

Surveys

The MLQ is only one of the many surveys that are used in research and practice to attain measures of leadership attitudes in business and studies (Arnulf, Larsen, Martinsen, & Bong, 2014). The MLQ, like many other leadership surveys, has a format that allows the respondents to choose on a five-point Likert-scale when they answer questions about a leadership situation. In this format, when a respondent has chosen the values she/he wishes to give, the answers are coded into a statistical measurement program that runs an analysis to give some meaning to the data. Usually when a measurement scale is to be validated there are requirements of internal consistency, often measured by Cronbach's alpha. This gives the researchers an indication of how well the items of the scale measures the same underlying construct. Also, exploratory and confirmatory factor analysis tests are conducted to ensure that these items do in fact belong in the same scale, and measure the same latent variable (Arnulf, et al., 2014). The general assumption then is that the scale is valid and reliable if these confirmatory measures come to such conclusion.

However, these statistical methods of scale creation have been criticized for over half a century (Coombs & Kao, 1960; Arnulf, et al., 2014; Maul, 2017). For example, in his study of surveys, Maul (2017) found that surveys made up of items that did not mean anything passed the test of both Cronbach's alpha and confirmatory factor analysis. He compared well-established surveys to surveys

with questions made up of nonsensical words. In his analysis, he found that the nonsense questions were almost just as valid as the well-established surveys, at least when validated through measures of internal consistency and confirmatory factor analysis. Since these scales consisted of items that were in fact meaningless, respondents could not have made any "understanding" of these. Therefore, one should question surveys made up of scales that have been validated through such statistical methods.

Further, factor analysis has been found to always produce a last factor, coined "social utility function" (Coombs and Kao, 1960). This factor signifies how the data is structured and the meaning of the items. The structure and wording of the items of a scale has later also been argued to affect measurement outcomes. It has been claimed that the way people answer on surveys is influenced by both the wording of the questions, and the similarities between them (Edwards, 2008). These crucial studies demonstrate that surveys may be flawed before we hand them out, both in the survey construction, as well as the planned statistical analysis and validation of the measurement tool.

Semantic Theory of Survey Responses

In their study of survey research, Arnulf and colleagues (2014) examined the semantic overlap in survey research. Through the use of text algorithms, they showed how respondents answer not to the item itself, but answer according to what is semantically expected of them. The text algorithms (MI and LSA) were able to predict responses to surveys that would be obtained from real human subjects and explained 60-86% of the variation in the sample. Traditionally, the variation found in the differing responses has been considered to be the effect or influence of a social or psychological variable. However, these results could indicate that the variance obtained from a survey reflect the semantic overlap between items within the scale. This has been demonstrated scientifically with other constructs within the organizational behavior (OB) field, and thereby the issue with semantic overlap among constructs seems to be a characteristic of the survey measurement tools themselves (Nimon, Shuck, & Zigarmi, 2016).

The Semantic Theory of Survey Response (STSR) proposes that when given a survey, respondents will see the similarity between items and will thereby try to be consistent in their answers (Arnulf, et al., 2014). Considering the

problematic semantic overlap in survey research, it becomes even more challenging as most leadership research is based on survey methods (Yukl, 2012, as cited in Arnulf & Larsen, 2015). The issue with such findings arrives when trying to measure attitude strength while these surveys in fact measures something entirely different, namely semantic consistency in responses (Arnulf, Larsen & Martinsen, 2018). The MLQ for example attempts to measure attitudes towards one's leader, but instead, the researcher ends up with data reflecting semantic structures of relatedness between items and concepts. In conclusion, it could therefore be expected that statistics generated from individuals answering survey questions, only reflect the semantic processing of the item instead of judgements/attitudes towards the leader.

To further illustrate Arnulf, Larsen, Martinsen, and Egeland (2018) explains Wittigenstein's Tractatus Logico-Philosophicus model in relation to semantics in research. To understand how answering surveys relate to semantic processing, Wittigenstein's presentation of different facts is a good example. There are three different types of facts: empirical facts, psychological facts, and logical facts. Empirical facts are the ones researchers are looking for when running a controlled experiment. These facts have been established and re-tested by scientific observation. Then there are psychological facts which are facts about what people *believe* to be true or not. It is different from empirical facts in that empirical facts have been tested and could be said to be universally true, while psychological facts only say something about what individuals believe to be true. Thirdly, the logical facts are a prerequisite for understanding the other two, and thus constitutes the comprehension of the statement. A statement must also be logically accepted and interpreted. One needs to be able to differentiate this statement with all other statements, to accept its truth.

According to Arnulf and colleagues (2018), the problem arises when researchers takes psychological facts (what people believe to be true) as empirical facts (scientifically observed to be true). This implies that researchers within the OB community take people's attitudes towards their leader as empirical research. Once again, in Arnulf's and colleagues most recent study, they showed that the patterns in their data could be explained by semantics. If the patterns in the data can be explained by semantics, they are not empirical facts, rather they are logical ones (Arnulf, et al., 2018). The patterns explain logical relationships of how people understand and are able to see the similarity between the items in a survey.

If this is true, then the two mechanisms (assigning attitudes vs. comprehending items) are of a different nature. Both of these could be considered psychological mechanisms, but Rensis Likert did not intend language comprehension to be the main process when answering items on a five-point Likert scale (Likert, 1932). It seems as though surveys are not able to clearly define the two. Theoretically, these two mechanisms can be traced neurobiologically in the brain in which one mechanism should be related to logical processing (language parsing processing) and the other should be related to emotional processing (assigning attitudes). We continue our review of the distinct neurological processes of semantics and emotions to prepare STSR for a potential fMRI pilot experiment.

Neuroscience of Semantics

When we use the term semantics, we are referring to the meanings of words which individuals come to know and understand through acquired knowledge and interactions. For example, things such as shapes, colors, sounds, movements, actions, and environments, we come to know and understand through our experiences. This type of knowledge is represented symbolically by language (Binder, Desai, Graves, & Conant, 2009). In other words, the semantics of a language are the relationships between words and our stored knowledge about the world. Semantic processing is the act of making associations while interpreting words and sentences. This differs from solely phonological or visual processing (Binder, et al., 2009) and therefore is more than simply reading letters or hearing sounds. Rather, semantic processing is making associations effortlessly and continuously.

How semantics and language processing take place in the brain has been under close examination within the last couple of decades. Binder and colleagues (2009) in their meta-analysis of neuroimaging studies of semantic processing identify seven brain regions of documented activity: posterior inferior parietal lobe, lateral temporal cortex, ventral temporal cortex, DMPFC (dorsal medial prefrontal cortex), IFG (inferior frontal gyrus), ventromedial prefrontal cortex, and posterior cingulate gyrus. The angular gyrus was the region that showed the highest activation foci in their study in terms of semantic processing. The angular gyrus is associated with retrieval of concepts and conceptual integration, which is

an important aspect of semantic processing (Binder, et al., 2009). Further, other fMRI studies have shown how this region is activated when individuals engage in sense-making of words (Newman, Just, Keller, Roth, & Carpenter, 2003). Also, Humphries, Binder, Medler, & Liebenthal (2007) showed in their study of sentence comprehension that subjects listening to words had a delayed activation in the angular gyrus, in relation to baseline of the other brain regions activated. This suggests that this region is also of importance to semantic processing.

The middle temporal gyrus (MTG) is another region consistently shown to be active during semantic processing (Dronkers, Wilkins, Van Valin, Redfern, & Jaeger, 2004) since damage to this brain region causes loss of language comprehension, and severe semantic deficits (Binder, et al., 2009). Further, the left dorsal medial prefrontal cortex (DMPFC) is important to the role of semantic retrieval (Binders, et al., 2009). Studies of patients with damage to this area have trouble producing sentences of meaning, but are able to repeat simple words (Robinson, Blair, & Cipolotti, 1998).

Another semantic region identified was the posterior cingulate gyrus. In Binder, et als. (2009) meta-analysis's, the posterior cingulate gyrus was most consistently activated when subjects engaged in semantic processing. It has however, been linked to several higher-order cognitive functions such as emotional processing, working memory, spatial attention, and visual imagery, just to mention a few. Binder and colleagues (2009) argue that the posterior cingulate gyrus may be important to episodic memory, and that the relation to semantic processing is how individuals store and retrieve information in their episodic memory. Another nearby brain region that has consistently been shown to activate in semantic processing, and other cognitive functioning is the rostral cingulate gyrus (Kuchinke, Jacobs, Grubich, Võ, Conrad, & Herrmann, 2005). Kuchinke and colleagues (2005) found that the rostral cingulate gyrus was activated when individuals processed emotional words.

In a recent study by Huth, Heer, Griffiths, Theunissen, and Gallant (2016), researchers created an algorithm (PrAGMATiC) that models an intricate system of semantic information across broad regions of the prefrontal cortex, lateral and ventral temporal cortex, and lateral and medial parietal cortex. In the study, subjects were scanned while listening to stories, totaling a 10,470 word-lexicon by completion. By categorizing and locating these words, researchers were able to

create what they called a single "atlas" of the cerebral cortex. Patterns appeared to be relatively consistent across individuals.

Based on these findings, during an fMRI study of survey responses, we would expect to see activation mainly in the posterior cingulate gyrus due to its role in conceptual processing. We would also expect supporting activity in areas such as middle temporal gyrus for its role in comprehension and a selective portion of the cerebral cortex for word retrieval.

Neuroscience of Emotions and Attitudes

Attitudes or "hot cognition" can be thought of as emotionally-laden judgements. These appraisals are considered the cognitive antecedent of emotion and are evoked by the evaluation of significance of circumstances for personal-wellbeing (Smith, Haynes, Lazarus, & Pope, 1993). As various cognitive appraisal theories would suggest, emotions are the product of evaluation outcomes believed to be either positive or negative. A positive evaluation of a circumstance leads one to a potential "benefit" outcome assumption, whereas a negative evaluation indicates a potential "harm" outcome assumption. Utilizing such theories, researchers Smith and Lazarus (1990) created a two-level model that breaks down the evaluation process of hot cognition. At the primary level, appraisal components are the actual specific questions to be evaluated in the appraisal process ("Does this affect me?"). The secondary level explores the core relational themes of the evaluation ("If so, in what way?"). The themes represent the patterns of the answers to the appraisal questions that have special significance ("Is this a benefit vs. harm?") (Smith & Lazarus, 1990).

Undeniably, emotions, attitudes, and emotionally-laden judgements involve complicated cognitive activity, but not *all* cognitive activity is relevant to emotion, and even relevant cognitive activities are not all *equally* relevant (Smith & Lazarus, 1990). Put forth in 1949 by MacLean, The Limbic System Theory of Emotion proposed that emotions exist within a specialized group of neural structures within the limbic system-- working collectively to form a singular system housing emotion (Murphy, Nimmo-Smith, & Lawrence, 2003). However, since those early days, we have come to know now that emotions stem from much more specific brain structures, as neuroimaging studies would suggest (e.g. Vytal & Hamann, 2010). In fact, what we know today is that many non-limbic system

structures can become activated simultaneously when it comes to emotions and emotion processing (Murphy, et al., 2003). For example, the ventromedial prefrontal cortex (VMPFC) has received much attention lately for its role in emotion-related activity, specifically for schematic processing systems. The schematic emotional processing system operates through integrating perceptual and sensory information typical of a given category of emotional experiences (Schaefer, Collette, Philippot, Van der Linden, Laureys, Delfiore, & Salmon, 2003). It is the subjective experience associated with schematic processing that initiates a spontaneous way of appraising a situation. We believe this is important to keep in mind when thinking of hot cognition in answering surveys about one's leader.

In a meta-analysis, researchers (Vytal & Hamann, 2010) used the Activation Likelihood Estimation Method (ALE Method) to analyze results of neuroimaging studies examining locations of emotional activity in the brain. The ALE method is unique in that it preserves the voxel "coordinates", instead of assigning regional labels to the activation coordinates. Through applying this method, researchers found significant associations between emotion states and regions of the brain. The consistent patterns of activation for the basic emotions follows: Happiness, right superior temporal gyrus; Sadness, left medial frontal gyrus; Anger, left inferior frontal gyrus; Fear, left amygdala; Disgust, right insula and right inferior frontal gyrus (Vytal & Hamann 2010). These areas would activate when an emotionally-related response or judgement is taking place.

Predicting, identifying, and distinguishing emotions from neuroimaging data is quite challenging, as brain regions often participate in several emotions at once, or share other cognitive processes at the same time. Hypothetically, if emotional-ladened judgement is taking place while answering surveys, versus just semantic, language parsing processing, we would expect to see stronger activation of neural networks-- potentially within the VMPFC or the above-mentioned brain regions.

Current fMRI Methods

Using a developmental approach to test STSR neuroscientifically, we conducted the above described fMRI literature review and met with fMRI technical professionals to discuss potential experimental designs. We also

underwent short training sessions in the method. However, it should be noted that fMRI is a highly complicated technique, suitable for more simplistic experimental designs. During fMRI scans, there is a lot of noise that appears in the signals which makes many variables hard to control for. Since we are looking at several systems -- emotional processing systems and semantic processing systems, a significant level of noise is generated. Both systems involve brain structures and locations that are distributed across the cerebral cortex, making it difficult to focus and parcel out any one or two unique processes (Kanwisher, 2010) that we can confidently point to as providing an answer to our burning question.

We came across some other challenges, which were also noted by Hauk & Tschentscher (2013). These researchers stated that there are three challenges that fMRI needs to overcome in order to start looking at semantics and systems more seriously: 1)Activation is correlational, and may not be causally related to the processes of interest; 2) Activation can be ambiguous with respect to the processing stage at which it occurs; and 3) There is no one-to-one relationship between brain areas and cognitive functions (2013). The described difficulties were to be expected and have been carefully considered, resulting in the need for more specific experimental design features. The next step is to test STSR by continuing our exploration of this research question with a more behavioral approach, in which we will try to manipulate both the logical and psychological aspects of Wittigenstein's model to prepare for the empirical studies of an fMRI.

Behavioral Methods of Testing STSR

Before we can uncover the underlying cognitive processes to further examine the Semantic Theory of Survey Responses, we look to other ways to test more indirectly. At this point, we should first address the controversiality of STSR in the social science research community. Majority of leadership research has up until now greatly depended on surveys with Likert scales as the main way to gather and analyze information on attitudes (taken to be truths), and STSR questions the very validity of it all. STSR takes away the tool of the social scientist, in a sense, and may in fact invalidate much of the already acquired data. It leaves many to beg the question that if surveys are just collecting semantic information, what can we do about it? If STSR is in fact valid, can we intervene in the survey response process in a way that elicits actual assessment on the part of

the participant, rather than producing semantically predictable responses? Specifically, it has been suggested that a manipulation of temporal characteristics of survey administration, may in fact produce responses that go beyond semantic expectation.

Manipulation of Temporal Characteristics

What is filling out a survey, other than making a series of quick decisions? Based on the work of Arnulf and colleagues (2018), the automatic language parsing process that takes over is the semantic processing of the questions themselves. This is a fairly unconscious, intuitive process that takes little time or effort on the part of the survey participant.

In our experiment we aim to explore the response patterns of individuals who go through a time-controlled survey. Previous research suggest that people who are forced to take a pause before answering an item on a survey will give better quality answers (Kapelner & Chandler, 2010). In a series of survey studies using Amazon's MTurk database, researchers Kapelner and Chandler (2010) found that by implementing waiting periods for survey responses, the quality of survey responses increased by 10%. Researchers theorized that it was due to the extra time spent thinking and thoughtfully answering questions, instead of satisficing (settling for the easiest choice). Pausing before answering an item might nudge participants to consider the question beyond the meaning of words, and then generate a legitimate evaluation, which will lead to a more informed judgement or attitude. In this case, introducing a pause would interrupt the semantics, and yield a greater evaluation or response.

However, research in this field strongly suggests that surveys are constructed in a way that makes them semantically dependent, thus changing the response time will have little impact on how people answer. As we have already discussed, when people answer surveys in organizational psychology and leadership they answer what is semantically expected of them. It could therefore be that surveys are created in a way that makes them semantically predictable even before the surveys are given out. In fact, according to Arnulf and colleagues (2018) the best way to make sure a survey passes the goodness of fit test is by creating it based on mere semantics. Therefore, the responses from the participant surveys will still be predictable using the semantic algorithms. We anticipate that

we will see semantic variation that will be predictable by semantic algorithms despite a temporal manipulation of survey administration.

In keeping with a developmental approach to piloting an fMRI study, introducing a pause to the experiment will help us in two ways. First, a temporal manipulation will allow us to behaviorally test STSR to examine how semantics are affected via survey response analysis. Secondly, a forced pause will serve as an experimental design feature that could be re-produced in the fMRI lab later. By allowing more time between the question and response, the fMRI technique may be able to better detect and differentiate between cognitive processes occurring, mainly semantic vs emotional processes.

To begin our behavioral study, we will use Kapelner and Chandler's same forced-pause method to examine whether or not responses show more semantic differentiation or predictability than previous traditional survey administration methods.

Methods

Participants

Our sample consisted of 89 (34% male, 66% female) survey subjects, with majority of respondents (64%) within the age range of 25-34. 62% were employed full time, 9% were employed part-time, 25% indicated they were students, 3% stated current self-employment, and 1% indicated they were retired. Recruitment for the study was conducted through convenience sampling. An anonymous link was posted on LinkedIn and Facebook, so people who were interested in taking the survey could click on the survey link. The description of the survey was presented as a leadership survey where respondents were asked to answer statements about a leader/supervisor they had before, or a current leader.

Experiment

Due to MLQ's ubiquitous use in leadership research and development, we will use this questionnaire--specifically the transformational leadership scales--for our survey. Transformational leadership is measured by 5 scales, totaling 20 items (see Appendix A for list of questions). Subscales include Idealized Influence--both attributes and behaviors (building trust, inspiring power and pride),

Individualized Consideration (personalized coaching, attending to employee needs), Intellectual stimulation (inspiring creativity and encouraging intellectual pursuits), and Inspirational Motivation/Leadership (articulating shared goals and encourages mutual understanding of purpose) (Bass & Avolio, 1993). By using only the transformational leadership scale, we can reduce the time spent on the survey for our respondents to ensure better completion/less attrition during the study. Since long and complex measurement technique often results in elevated dropout rates (Fagarasanu & Kumar, 2002), using the transformational scale will give us the most relevant and complete information.

The survey was created in Qualtrics, which is provided by BI. Qualtrics is an online survey tool that allows users to design their own survey and generates a link for participants. The survey was designed to give respondents 15 seconds to think about the item/statement before answering. To begin, the respondents were first given an introduction which stated that the survey intended to measure the individual respondent's leader, and the respondent was given assurance that responses were confidential and anonymous. Respondents were then instructed to continue to the first item. The item was presented together with a timed clock counting down the 15 seconds. When the 15 seconds were over, they were guided through to the next page which gave them the original answering options of the MLQ ranging from: Not at all - Once in a while - Sometimes - Fairly often - Frequently, if not always. The respondent had to choose which of these options suited them best, and had to click on the next button themselves. This design continued for all 20 items (see Appendix B for survey format examples).

Once all the MLQ items were answered, the respondents were asked to indicate their age, gender, and occupational status. Next all respondents were given contact information to the researchers in case they would have any questions.

Total number of subjects that completed the entire survey was 92. Together with this, a sample was provided to us to compare with the same 20 questions answered by 92 different subjects that completed the survey without a pause. The two groups were considered to consist of mostly native English speakers.

Analysis

The method for data analysis used in this experiment has been adopted from Arnulf and colleagues (2018). To begin our analysis, we start by first creating an "item distances response matrix", which will tell us how similar or dissimilar two items are. The similarity between two items is the attributed semantic similarity. Creating an item distance matrix will theoretically help us to decouple the semantic influence from attitudinal or emotional influence in a person's response because we are only calculating the absolute difference between the responses, while ignoring attitude strength. Next, we will then compare our item distances to semantic algorithms (MI & LSA) we run on the responses to see how closely these numbers relate. In addition to the method of analysis outlined by Arnulf et al. (2018), we continue our exploration via factor analysis and further correlation analysis.

Item-distances (190 item distance pairs)

First, items-distance pairs were created for all participants. The item-distance matrix uses the similarity between two items to give an estimation of the likelihood that these items may be similar. An item-distance is calculated by subtracting one item with another, for example item1 with item2. If a respondent answers 5 on item1 and 4 on item2, then the item-distance is 1 (5-4=item-distance). The rationale is that when item distance is low, then the two items are perceived to be similar. However, when item distance is high, the two items are perceived to be very different. These values do not contain information about attitudes, as the attitude strength is removed in this process. By using item distances we get an indication of whether the respondents perceive the items to be similar. If the items are perceived to be very similar, then it is likely that answering one will be indirectly answering the second. We do not know why these items may be perceived to be similar, which is why we need language parsing algorithms.

Language parsing algorithms

MI is a language algorithm that looks at sentence similarity using a corpus-based approach. The MI algorithm bulks the words from the two sentences

it is comparing into word classes. Then all the words from sentence 1 are compared to all the words in sentence 2 using the word classes as a way of classifying which are to be compared into WordNet hierarchies. Lastly, the highest semantic similarity for each word is normalized using the "inverse document frequency" from the British National Corpus for the weighting of uncommon and rare words. This will eventually give us not only the degree of word similarity, but the semantic similarity of the two sentences (overlap of meaning). This technique was used to compute the similarity between our 20 MLQ items.

The LSA algorithm uses another kind of semantic information than the MI algorithm. LSA is short for Latent Semantic Analysis and it focuses on the co-occurrence of words in the same text. While the MI looks for word classes, LSA looks for latent variables in the similarity of words and how they co-occur. If we take "Red" and "Merlot" these words are not usually connected, but they will be in a text if the word "Wine" occurs in the same document. PRTheory, NewsTheory, BIZtheory which will be shown in the further analysis, are names for the same algorithm that weights the semantic spaces differently. Therefore, when analysing the data, the different algorithms (MI and LSA) will look for different semantic similarity and will weight information differently.

Item-distance groups

The algorithms that are used in this experiment have analysed the items in the MLQ and also created pairs which indicate item-item similarity. This means that we are left with three groups of item pairs. Each of these three groups contains 190 item-distance pairs in our experiment. The first two groups come from respondents, the experimental and control group. These two are what we refer to throughout the paper as the item-distance pairs. The third and last group of item-distances we are left with has been created as a result of the language algorithms identifying how much semantic overlap there is between items in the survey. The plan is then to compare these three groups on the individual "item-pair" level to detect semantic patterns.

Item distance analysis

Initially, the item-distances were correlated with the MI values for the same item pairs to explore whether there was a semantic relationship in the data or not. It should here be noted that the MI operates with numbers between 0 and 1, and that a high number means a high semantic relationship. However, the item distances operates with numbers between 0 and 4, with a low number indicating a relationship between the two items. Therefore, these will be inversely related. This means that the higher the semantic similarity between two items, the higher the value MI will be. However, when the item-distance is low in the response groups, there is a high similarity. This is the case with LSA also. We ran a regression analysis on all the individual item-distance pairs.

In order to find how much of the semantic algorithms could explain variation in our sample, a regression analysis was applied for each of the 92 respondents. The MI algorithm, and the LSA algorithm were used as prediction variables. This analysis gave us the adjusted R-squared values for the entire sample. The same procedure was applied to the 92 respondents from the control group. These tests would give an indication of how much of the responses can be attributed to semantic structures.

The resulting adjusted R-squared values from the experimental group and the control group were compared in an independent-samples t-test in SPSS. This test was performed to see whether the two groups significantly differed in terms of their semantic predictability as were hypothesized.

Earlier studies using this method have considered MI to be the easiest to work with in shorter surveys (Arnulf, et al., 2014; Arnulf & Larsen, 2015; Arnulf, et al., 2018). Arnulf and colleagues (2018) argued that using both would be better at predicting natural speech, and we have therefore decided to use them both in our analysis. However, because of the differences between the MI algorithm and LSA algorithms in how they process text, we decided to run them together and separately. This would give us an indication of how much the different semantic algorithms contributes to variation in our sample. Running them separately and together will also give valuable insights for future studies using the MI and LSA as tools for analyzing survey items.

Factor analysis

According to research theory, when conducting factor analysis on transformational leadership questions on the MLQ, there should be 4-5 factors, considering the concepts of Individualized Consideration, Intellectual Stimulation, Inspirational Motivation, Idealized Influence Attributes and Idealized Influence Behaviors (Bass & Avolio, 1993). However, this does not always occur, and so the MLQ is frequently criticized for its structural validity. In fact, in a study done by Tepper and Percy (1994), the MLQ's latent structure of two independent samples was examined through confirmatory factor analysis and it was found that Idealized Influence and Inspirational motivation scales converged to form a single latent construct. Work later done by Carless (1998) further questioned the 5-factor transformational leadership model, stating that the MLQ does not measure separate transformational leadership behaviors but instead appears to assess a single construct. Since then, many researchers have questioned the factor-model (Leivens, Van Geit, & Coetsier, 1997; Tejeda, Scandura, and Pillai, 2001; Tracey and Hinkin, 1998; ; Tracey, Hinkin & Enz, 1997)

To identify some of the differences between the control group and experimental group, exploratory factor analysis was conducted. Factor analysis gives additional information of how people think about the questions in terms of how they should cluster together. We decided to run an exploratory factor analysis on each separate group, as well as on the experimental and control groups together. If the groups are similar, then the factor analysis from both groups should show the same clustering of items.

Correlation analysis

In an attempt to explore the factor structure with additional analysis, correlations were computed between the 5 different dimensions of Transformational Leadership and the individual factor loadings from each individual. The factors and each participants' contribution to the factor loadings was saved from the initial exploratory factor analysis and correlated against the average scores from the different Transformational Leadership dimensions. This way we can explore the relationship between the experimental and control group in terms of how they well they fit with the dimensions. Because the initial exploratory factor analysis showed that the experiment group loadings were

distributed among 5 separate factors while the control group only appeared to load on 2 factors, these two solutions were the ones we used to explore whether the factor loadings from the initial exploratory factor analysis were matching the theory of the five dimensions of TL.

Additional correlations were computed between the semantic algorithms and the 5 TL dimensions. The correlations of the semantic algorithms will help us explore the relationship between the five dimensions and semantics, and whether the dimensions are semantically predetermined or not.

Results

Average scores on the 5 dimensions of Transformational Leadership from the two groups (control vs. experimental) are outlined below in table 1. What we found was that the experimental group had higher means on almost all dimensions except for Idealized Influence-Behavior. We also see a trend of lower variation in the experimental group, so we can infer that respondents could be answering more similarly to each other.

Table 1:Descriptives and t-test results of the 5 Transformational Leadership dimensions in the control and experiment groups

IIA (Idealized Influence Attributes)	Mean	Std.Deviation	N	F	Sig.
CONTROL	3,34	1,02	92		
EXPERIMENTAL	3,44	0,83	92	5,67	.01*
IIB (Idealized Influence Behavior)	Mean	Std.Deviation	N	F	Sig.
CONTROL	3,38	0,99	92		
EXPERIMENTAL	3,24	0,79	92	4,16	.04*
IM (Inspirational Motivation)	Mean	Std.Deviation	N	F	Sig.
CONTROL	3,45	1,02	92		
EXPERIMENTAL	3,77	0,69	92	17,23	.00**
IS (Intellectual Stimulation)	Mean	Std.Deviation	N	F	Sig.
CONTROL	3,28	0,91	92		
EXPERIMENTAL	3,40	0,81	92	2,33	.12
IC (Individual Consideration)	Mean	Std.Deviation	N	F	Sig.
CONTROL	3,18	0,99	92		
EXPERIMENTAL	3,56	0,88	92	0,76	.38

*$p<.05$. **$p<.01$

T-test showed that the experiment and control group differed significantly in their responses on three out of the five dimensions, namely, IIA ($F(183) = 5.67$, $p = .01$), IIB ($F(183) = 4.16$, $p = .04$), and IM ($F(183) = 17.23$, $p = .00$). This tells us that the temporal manipulation had some effect.

The regression analysis with the item-distance pairs and the semantic algorithms (MI, PRTheory, NewsTheory, BIZTheory) showed that the semantic algorithms explained about 6% (the lowest amount considered significant) to 42%

of the individual variation in our experimental group, and 6%-49% of the variation in the control group (see Table 2).

Table 2 shows the distribution (Std. Deviation) of the adjusted R-squares taken from the regression analysis of the item-distance pairs and semantic algorithms. The experimental group has a higher mean (0,06) and more variation with a Std. Deviation of 0,08. This is further illustrated in the spread of R-squares (see figure 1 for experimental group, and figure 2 for control group).

Table 2: Mean, Std. Deviation, Min, Max from Adjusted R2

	Control	Experiment
N	92	92
Mean	0,05	0,06
Std. Deviation	0,07	0,08
Minimum	-0,02	-0,02
Maximum	0,49	0,42

The independent t-test showed that the experimental and control group did not significantly differ from each other (F (1, 179) = .93, p = .33) (See Table 3).

Table 3: Results of t-test and Descriptive Statistics for Semantic Algorithms by Group

	Group						95% CI for Mean Difference		
	Control			Experiment					
	Mean	SD	n	Mean	SD	n		t	df
Semantic Algorithms	0,05	0,07	92	0,06	0,08	92	-0,029, 0,0174	-4,94	182

Figure 1: Distribution of Adjusted R-Squared from the Experimental Group

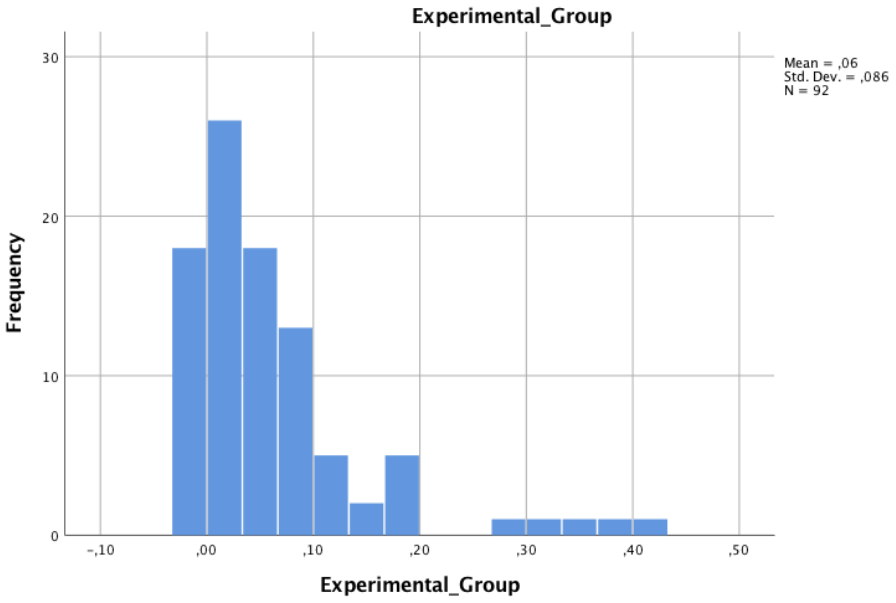
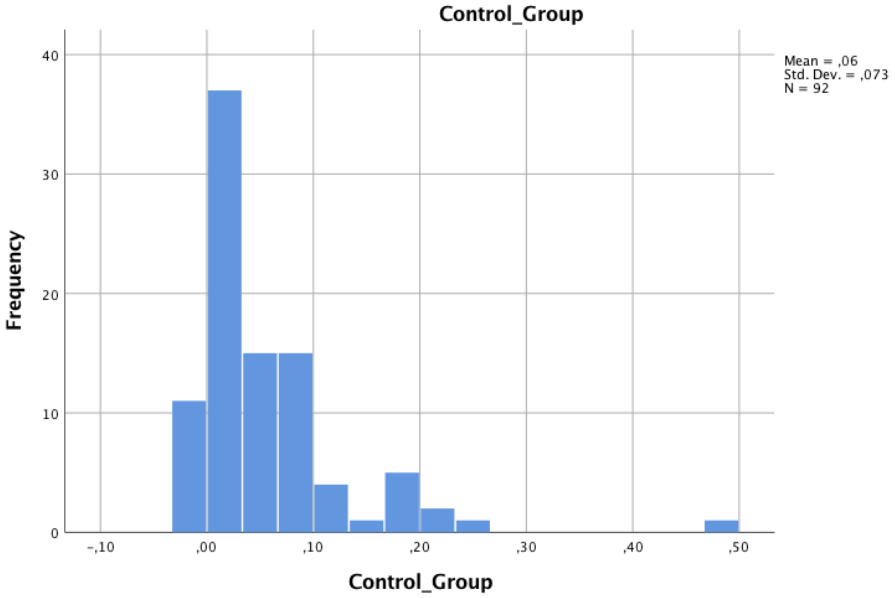


Figure 2: Distribution of Adjusted R-Squared from the Control Group



There was a difference in the explained variance when using the two different types of algorithms. The initial belief was that MI would yield a better result in terms of how much the algorithms would explain the variance, at least compared to the LSA. However, regression analysis on the separate algorithms (MI vs. LSA) shows that LSA is better at predicting semantic similarity in this sample, and also shows more variation (control - LSA: 41,2% explained variance, experimental - LSA: 33% versus control - MI: 11,2%, experiment - MI: 9,2%).

Using the principal component analysis method with varimax rotation, the factor analysis revealed that the two groups combined (n=184) loaded on 3 factors explaining 60,86% variation. Separately, the control group (n=92) factor structure loaded on 2 factors, with just one factor explaining 36.28% of the variation, for a total of two factors explaining 66.42% of the variation.The factor analysis that was run on the experimental group (n=92) shows that the sample loads on 5 different factors, explaining a total of 63.47% of the variation.

Table 4a: Rotated Components Matrix of Factorial Analysis of Control group

	Components	
	Factor 1	Factor 2
MLQ 2	**0,58**	0,37
MLQ 6	0,05	**0,84**
MLQ 8	0,47	**0,49**
MLQ 9	0,29	**0,77**
MLQ 10	**0,76**	0,35
MLQ 13	0,48	**0,72**
MLQ 14	0,41	**0,74**
MLQ 15	**0,68**	0,43
MLQ 18	**0,79**	0,32
MLQ 19	**0,80**	0,03
MLQ 21	**0,79**	0,40
MLQ 23	**0,65**	0,40
MLQ 25	0,37	**0,67**
MLQ 26	0,36	**0,84**
MLQ 29	**0,65**	0,30
MLQ 30	**0,70**	0,35
MLQ 31	**0,74**	0,38
MLQ 32	**0,77**	0,39
MLQ 34	0,50	**0,72**
MLQ 36	0,54	**0,62**

Note Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
Rotation converged in 3 iterations. Total explained variance: 66,42%

Table 4b: Rotated Components Matrix of Factorial Analysis of Experiment group

	Components				
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
MLQ 2	0,03	**0,73**	0,20	0,05	-0,03
MLQ 6	0,29	0,09	**0,72**	-0,12	0,06
MLQ 8	**0,61**	0,28	0,07	0,18	0,44
MLQ 9	0,13	0,06	0,43	0,31	**0,45**
MLQ 10	**0,71**	-0,19	0,24	0,04	0,07
MLQ 13	0,12	0,17	0,13	0,10	**0,74**
MLQ 14	0,05	0,25	**0,73**	0,24	0,06
MLQ 15	0,13	**0,66**	0,14	0,11	0,38
MLQ 18	**0,66**	0,44	0,08	0,18	-0,08
MLQ 19	**0,71**	0,02	0,15	0,07	-0,01
MLQ 21	**0,57**	0,41	0,20	0,36	0,06
MLQ 23	0,39	**0,44**	0,09	0,22	-0,54
MLQ 25	0,12	0,17	0,08	**0,79**	-0,11
MLQ 26	0,12	0,33	0,43	**0,47**	0,27
MLQ 29	**0,55**	0,38	0,46	0,05	-0,16
MLQ 30	0,53	**0,55**	0,10	0,20	0,22
MLQ 31	**0,57**	0,53	0,28	0,09	0,14
MLQ 32	**0,55**	0,27	0,02	0,34	0,28
MLQ 34	0,26	0,12	**0,62**	0,40	0,15
MLQ 36	0,17	-0,02	0,13	**0,74**	0,24

Note Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
Rotation converged in 9 iterations. Total explained variance: 63,47%

After suppressing small coefficients (<.40), based on the rotated component matrix, there are cross loadings in both the control and experimental group. In the control group, there were greater differences between loadings,

whereas in our experimental group, several of the cross-loadings were less than or equal to .20 differences (See table 4A & 4B for Rotated Component Matrices). Nevertheless, there is a clear 5 factor structure of the experimental sample, compared to the 2 of the control group.

Table 5: Correlation matrix for the forced factor solutions for the control group, and semantic algortihms (N=92)

Variables	1. Regression Score Factor 1 (Control)	2. Regression Score Factor 2 (Control)
1. Intellectual Stimulation (IS)	0,77**	0,48**
2. Inspirational Motivation (IM)	0,46**	0,83**
3. Individual Consideration (IC)	0,88**	0,34**
4. Individualized Influence Behavior (IIB)	0,48**	0,81**
5. Individualized Influece Attributes (IIA)	0,80**	0,50**

**p<.05. **p<.01.*

After the exploratory factor analysis revealed that the experiment group loaded on 5 factors like the original 5 TL dimensions, while the control group only showed loadings on 2 factors, we used a forced-factor structure method that gave us how much each participant contributed to this particular structure (both experiment and control group). This information was then used to explore how the 5-factor solution and the 2-factor was distributed in relation to the TL dimensions.

The 2-factor solution from the control group showed that the factors are significantly correlated with the TL dimensions where Factor 1 revealed correlations from 0,46 up to 0,88, and Factor 2 with correlations from 0,34 to 0,81(Table 5). Since all dimensions were significantly related to the two factors, we cannot attribute a specific TL dimension to any factors. Also, this factor solution that was created for the control group was not significantly related to any of the language algorithms based on the correlation analyses.

Table 6: Correlation matrix for the forced factor solutions in the experiment group (N=92)

Variables	1. Regression Score Factor 1 (Experiment)	2. Regression Score Factor 2 (Experiment)	3. Regression Score Factor 3 (Experiment)	4. Regression Score Factor 4 (Experiment)	5. Regression Score Factor 5 (Experiment)
1. Intellectual Stimulation (IS)	0,45**	0,58**	0,13	0,28**	0,48**
2. Inspirational Motivation (IM)	0,08	0,07	0,40**	0,56**	0,64**
3. Individual Consideration (IC)	0,51**	0,57**	0,34**	0,13	0,29**
4. Individualized Influence Behavior (IIB)	0,27**	0,35**	0,80**	0,29**	0,01
5. Individualized Influece Attributes (IIA)	0,62**	0,38**	0,20*	0,49**	0,10

**p<.05. **p<.01.*

The 5-factor solution that was revealed in the experiment group showed a better differentiation and a clearer pictured emerged. Factor 1 was significantly correlated with 4 of the TL dimensions (IntStim, IndCon, IdeaBeh, and IdeaAtt) with correlations reaching 0,62 for Idealized Influence Attributes. Factor 2 was also significantly correlated with 4 dimensions (IntStim, IndCon, IdeaBeh, and IdeaAtt) where the highest correlation was 0,58 for Intellectual stimulation. Factor

3 was significantly related to 4 dimensions as well (InspMot, IndCon, IdeaBeh, and IdeaAtt), with the highest correlation being Idealized Influence Behavior at 0,80. Factor 4 was also significantly related to 4 of the TL dimensions (IntStim, InspMot, IdeaBeh, and IdeaAtt), the highest being Inspirational Motivation at 0,56. Lastly, Factor 5 was significantly correlated to 3 of the 5 (IntStim, InspMot, and IndCon) TL dimensions where the highest was Inspirational Motivation at 0,64. (Table 6)

Table 7: Correlation Matrix between the semantic algorithms and the TL factor structure

	MI	PR	News	BIZ
1. Intellectual Stimulation (IS)	-0,19	-0,22*	-0,05	-0,17
2. Inspirational Motivation (IM)	-0,04	-0,29**	-0,15	-0,33**
3. Individual Consideration (IC)	-0,18	-0,24*	-0,00	-0,16
4. Individualized Influence Behavior (IIB)	-0,12	-0,30**	0,16	-0,20*
5. Individualized Influece Attributes (IIA)	-0,31**	-0,30**	-0,29**	-0,34**

*$p<.05$. **$p<.01$.

Correlation matrix of the 5-factor structure against the semantic algorithms revealed that some of the TL dimensions were significantly related to semantics. It is important to keep in mind that the algorithms and TL dimensions are inversely related to each other, and that a negative correlation is what is expected in this case. The MI was significantly related (-0,31) to Idealized Influence Attributes, but none of the other dimensions. The LSA algorithm PR was significantly related to all the TL dimensions (IntStim: -0,22, InspMot: -0,29, IndCon: -0,24, IdeaBeh: -0,30, IdeaAtt: -0,30). NEWS was significantly correlated with Idealized Influence Attributes with -0,29. BIZ was significantly related to InspMot (-0,33), IdeaBeh (-0,20), and IdeaAtt (-0,34). (see Table 7)

Discussion

The purpose of this study was to further explore the Semantic Theory of Survey Responses (Arnulf, 2014) in two different ways. First, we sought to find a way to test the connection between semantic vs emotional processing within the brain, either by neuroscience or behavioral studies. Secondly, we wanted to explore the possibility of interrupting automatic semantic processes by manipulating survey administration. Throughout this process we have encountered several ideas for next steps in the development and testing of the theory.

The results of our experiment tell us several things. First of all, the results from the general experimental vs control group t-test indicate a statistically

significant, though small, variation between the two groups. This means that the temporal manipulation of the MLQ has at least some effect.

When testing through the use of regression analysis using the R-squares, we see that semantics play a role in the responses and that some of the variation in the sample can be explained by the semantic structures. However, the two samples do not differ significantly in semantic predictability using the MI and LSA algorithms, though the LSA-type algorithms performed slightly better.

It seems as though that despite the temporal manipulation creating a small significant difference between our experimental and control group, about 6%-40% of variation in both groups are attributed to the semantics of the questions themselves. Furthermore, because of the 2 factor loadings in the control group versus 5 factor loadings in the experimental group, we know that there are distinct differences between the methods of survey administration.

The factor analysis tells us that given extra time, people could potentially be responding to nuances in the semantics of the questions, which are undetectable by the regression and correlation analysis. It could indicate that people really are just answering to semantics--meaning that respondents were only reading, comprehending and trying to group the questions and be consistent in their answering style - just like STSR predicts. The fact that we were able to get the 5-factor structure on such a small sample is also impressive, and indicates categorization on the part of the participant.

It is important to re-emphasize at this time that when researchers use the MLQ Transformational Leadership scale they are looking for the 5-factor structure, but do not always come this solution. As mentioned previously, this is a contentious point for leadership researchers because it casts doubt on the foundations of Transformational Leadership as a concept. The prevalent 5-factor structure in our experimental group could suggest that we are seeing a clearer sign of the connections between answering a survey and mental models. We suggest that people become more aware of the meaning of the items and perhaps become better at detecting slight differences and grouping the five dimensions together. It could be that when individuals are forced to pause there is a stronger and deeper semantic processing activated, leading to a similar result such as researchers Kapelner and Chandler (2010) found with their improved quality survey responses. The question is then if the process that is enhanced is the precision of providing a judgement/attitude or an improved language parsing process.

The latter of those two theories seems to be more likely. In order to consider a survey like MLQ a reliable and useful tool, it must demonstrate content validity. From this point of view, surveys are circularly flawed in that they are created, tested, and established as valid in a way that can only be attributed to semantic similarities. Not only does the statistical analysis pick up on this, but so do the participants, as they were spending time on our survey and responding.

It is also quite likely that the Likert scale is to blame when it comes to poor survey construction and the predictability of semantic responses. The Likert scale may not be appropriate to tell us anything about the quality of a leader as it is originally intended. The scale is not designed to give differential answers, but rather detects semantic relationships due to the generally negative or positive responses that participants often give. Decades ago, in a study on Likert scales and quality of life evaluations, researcher Peabody (1962) was conducted analysis on survey responses by first calculating the relative contribution of response direction (negative or positive) and the response intensity (distance from neutral point). He found that the intensity of the evaluation only contributed 10 percent to the composite variation. This further supports the idea that the Likert scale, especially a 5-point scale, is just not sensitive enough to pick up anything more than general semantic differences, rather than degrees of attitude. We, as a collective research community, know this, however, the Likert scales continue to be used.

To take the uncertainty of method bias in survey construction and survey statistical analysis out of the leadership and semantics research, the need for advanced neuroscientific methods is evermore necessary. While single word and numerical cognition are well focused areas in semantics with a large body of neuroscientific and behavioral evidence (Hauk & Tschentscher, 2013), there is still a disconnect in process, representation, and causality. We need to find a way to take out the basic semantics from the true evaluation or attitude-- to turn the psychological and logical into the empirical. This is no simple task. However, recent breakthroughs in language neuroscience methods are increasingly becoming more specific. As previously mentioned, the research of Huth et al. (2016) consists of systematically mapping semantic selectivity across the cortex, essentially creating a word atlas. Future repetitions of this type of experiment could look into identifying other regions of leadership and leadership related words, while at the same time trying to identify other areas of significant neural

activity. This could be an interesting idea to pursue—finding out whether what we have "labeled" as leadership in our minds is part of an emotional response, or perhaps some primitive, yet evolving, part of the brain. In using our current study alongside Huth's method, one could see how closely grouped in the brain certain transformational leadership related words are and what other word networks are involved when asked leadership related questions. This could help link the behavioral response of categorization--what was apparent in our factor analysis--to the underlying related cognitive processes. However, these future research ideas assume that the fMRI techniques could be refined enough to process whole sentences or questions. They are also based on the corrections of the issues earlier stated by Hauk & Tschentscher (2013). Bringing fMRI or other future advanced forms of neuroscientific methods into the picture could help us understand leadership processes based on pure neurobiological facts rather than guessing about what the participant behavioral response score levels means.

Limitations

Due to the relatively new concept and method of this type of semantic research, our study is not without its limitations. Though the Semantic Theory of Survey Responses is still being explored and refined, we think it is important to discuss possible improvements for future research. Two of the main concerns involve subject perception and size limitations.

After completing our survey, we heard back from several respondents about their experience during the survey. Using the timer-feature of Qualtrics necessitates the appearance of a countdown clock in the upper corner of the screen. Several people provided feedback along the lines of feeling stressed or eventually annoyed with the forced wait time. Although we believe from this feedback that the timed response had the desired effect of increasing attention paid to survey questions--even annoyingly so--, in the future, the length of time could be tested and the timer hidden in the display of the survey. On the other hand, we also heard back from one respondent who mentioned the forced wait had a relaxing effect of them, surprisingly.

Since our sample size was small, we could not perform a confirmatory factor analysis, which we could have used to evaluate the factor structure within

the MLQ transformational leadership measurement model and determine how well the model fits the data. We would have been able to test the observed indicators of the latent variables of Individual Consideration, Inspirational Motivation, Idealized Influence Attributes & Behaviors, and Intellectual Stimulation.

Another issue that became apparent from our process was the unexpected performance of the MI vs the LSA semantic algorithms. Advances in language - parsing algorithms are happening daily and the semantic algorithms used (MI & LSA) may not be suitable to analyze this data in the future. The MI algorithm that has previously been thought of as the best tool for analyzing surveys like the MLQ proved to be worse at predicting semantic similarity than the LSA in our study. However, when combined they performed better. This could lead us to think that because of continuing advancements in language algorithms, there could be other algorithms that may be better at analyzing OB surveys. This should not be seen as a limitation, but rather that future research projects may be better off testing other algorithms.

Conclusion

The current study has contributed to the exploration of the Semantic Theory of Survey Responses by further testing its assumptions and preparing a potential fMRI project. With this study, we are advancing our understanding of the cognitive processes that are at play when answering survey items. We were able to show that even when people are forced to take a break and think about the question, the assumptions of STSR still holds. People are still only reading, comprehending and trying to be consistent when answering survey items. They are answering to what is semantically expected of them.

Leadership research has for a long time relied on these survey methods to get answers. Taking our study into consideration, along with other supporting evidence from Arnulf and colleagues (2014; 2015; 2018), the leadership field should question its methods. We are still not able to comprehend the complexity of this concept, perhaps because of the faulty methods we are relying on for answers. If leadership and motivation, along with other OB concepts are already highly correlated a priori because of their semantic relatedness, it does not serve

us any good in using surveys that will only confirm what we already know from before.

Using neurobiological measures to find answers may be the only way to test whether leadership does in fact exist, and how people actually feel about their leaders. This study has helped advance the understanding with regards to using fMRI as a method by going through a review of the latest research in the field, and providing a supporting behavioral experiment. Together, these pave the way for a future fMRI experiment to take place.

References

Arnulf, J. K., & Larsen, K. R. (2015). Overlapping semantics of leadership and heroism: Expectations of omnipotence, identification with ideal leaders and disappointment in real managers. *Scandinavian Psychologist*.

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., & Bong, C. H. (2014). Predicting survey responses: How and why semantics shape survey statistics on organizational behaviour. *PloS one*, *9*(9), e106361.

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., & Egeland, T. (2018). The failing measurement of attitudes: How semantic determinants of individual survey responses come to replace measures of attitude strength. *Behavior research methods*, 1-21.

Avolio, B. J., & Hannah, S. T. (2008). Developmental readiness: Accelerating leader development. *Consulting Psychology Journal: Practice and Research*, *60*(4), 331.

Bass, B. M., & Avolio, B. J. (1990). *Transformational leadership development: Manual for the multifactor leadership questionnaire*. Consulting Psychologists Press.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767-2796.

Carless, S. A. (1998). Assessing the discriminant validity of transformational leader behaviour as measured by the MLQ. *Journal of Occupational and Organizational Psychology*, *71*(4), 353-358.

Coombs, C. H., & Kao, R. C. (1960) On a connection between factor analysis and multidimensional unfolding. *Psychometrika*, 25: 219–231.

Demonet, J. F., Chollet, F., Ramsay, S., Cardebat, D., Nespoulous, J. L., Wise, R., Rascal, A., & Frackowiak, R. (1992). The anatomy of phonological and semantic processing in normal subjects. *Brain*, *115*(6), 1753-1768.

Dronkers, N. F., Wilkins, D. P., Van Valin, R. D., Redfern, B. B., & Jaeger, J. J. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition*, *92*(1), 145-177.

Eddy, P. L., & VanDerLinden, K. E. (2006). Emerging definitions of leadership in higher education: New visions of leadership or same old "hero" leader?. *Community College Review*, *34*(1), 5-26.

Edwards, J. R. (2008). To prosper, organizational psychology should… overcome methodological barriers to progress. *Journal of Organizational Behavior*, *29*(4), 469-491.

Evans, J. (2010). Intuition and Reasoning: A Dual-Process Perspective. *Psychological Inquiry, 21*(4), 313-326. Retrieved from http://www.jstor.org.ezproxy.library.bi.no/stable/25767204

Fagarasanu, M., & Kumar, S. (2002). Measurement instruments and data collection: a consideration of constructs and biases in ergonomics research. *International journal of industrial ergonomics*, *30*(6), 355-369.

Harrison, D. A., & McLaughlin, M. E. (1993). Cognitive processes in self-report responses: tests of item context effects in work attitude measures. *Journal of Applied Psychology*, *78*(1), 129.

Hauk, O., & Tschentscher, N. (2013). The body of evidence: what can neuroscience tell us about embodied semantics?. *Frontiers in psychology*, *4*, 50.

Hinkin, T. R., Tracey, J. B., & Enz, C. A. (1997). Scale construction: Developing reliable and valid measurement instruments. *Journal of Hospitality & Tourism Research*, *21*(1), 100-120.

Humphries, C., Binder, J. R., Medler, D. A., & Liebenthal, E. (2007). Time course of semantic processes during sentence comprehension: an fMRI study. *Neuroimage*, *36*(3), 924-932.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453-458.

Kanwisher, N. (2010). Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*, *107*(25), 11163-11170.

Kapelner, A., & Chandler, D. (2010, October). Preventing Satisficing in online surveys. In *Proceedings of*.

Kuchinke, L., Jacobs, A. M., Grubich, C., Võ, M. L. H., Conrad, M., & Herrmann, M. (2005). Incidental effects of emotional valence in single word processing: an fMRI study. *Neuroimage*, *28*(4), 1022-1032.

Lievens, P., Van Geit, P., Coetsier, F. (1997). Identification of transformational leadership qualities: An examination of potential biases. *European Journal of Work and Organizational Psychology*, *6*(4), 415-430.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Lunsford, L. G., & B. A. Brown (2017). "Preparing Leaders While Neglecting Leadership: An Analysis of US Collegiate Leadership Centers." *Journal of Leadership & Organizational Studies 24*(2): 261-277.

Maul, A. (2017). Rethinking Traditional Methods of Survey Validation. *Measurement: Interdisciplinary Research and Perspectives, 15*(2), 51-69.

Murphy, F. C., Nimmo-Smith, I. A. N., & Lawrence, A. D. (2003). Functional neuroanatomy of emotions: a meta-analysis. *Cognitive, Affective, & Behavioral Neuroscience*, *3*(3), 207-233.

Newman, S. D., Just, M. A., Keller, T. A., Roth, J., & Carpenter, P. A. (2003). Differential effects of syntactic and semantic processing on the subregions of Broca's area. *Cognitive Brain Research*, *16*(2), 297-307.

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

Peabody, D. (1962). Two components in bipolar scales: Direction and extremeness.*Psychological Review*, *69*, 65-73.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of applied psychology*, *88*(5), 879.

Podsakoff, P. M., & Organ, D. W. (1986). Self-reports in organizational research: Problems and prospects. *Journal of management*, *12*(4), 531-544.

Riggio, R. E. (2008). Leadership development: The current state and future expectations. *Consulting Psychology Journal: Practice and Research*, *60*(4), 383.

Robinson, G., Blair, J., & Cipolotti, L. (1998). Dynamic aphasia: an inability to select between competing verbal responses?. *Brain: a journal of neurology*, *121*(1), 77-89.

Rowold, J. (2005). Multifactor leadership questionnaire. Psychometric properties of the German translation Redwood City: Mind Garden.

Schaefer, A., Collette, F., Philippot, P., Van der Linden, M., Laureys, S., Delfiore, G., ... & Salmon, E. (2003). Neural correlates of "hot" and "cold" emotional processing: a multilevel approach to the functional anatomy of emotion. *Neuroimage*, *18*(4), 938-949.

Schedlitzki, D., Ahonen, P., Wankhade, P., Edwards, G., & Gaggiotti, H. (2017). Working with language: A refocused research agenda for cultural leadership studies. *International Journal of Management Reviews*, *19*(2), 237-257.

Smith, C. A., Haynes, K. N., Lazarus, R. S., & Pope, L. K. (1993). In search of the" hot" cognitions: attributions, appraisals, and their relation to emotion. *Journal of personality and social psychology*, *65*(5), 916.

Smith, C. A., & Lazarus, R. S. (1990). Emotion and adaptation. *Handbook of personality: Theory and research*, 609-637.

Tejeda, M. J., Scandura, T. A., & Pillai, R. (2001). The MLQ revisited: Psychometric properties and recommendations. *The Leadership Quarterly*, *12*(1), 31-52.

Tepper, B. J., & Percy, P. M. (1994). Structural validity of the multifactor leadership questionnaire. *Educational and Psychological Measurement*, *54*(3), 734-744.

Tracey, J. B., & Hinkin, T. R. (1998). Transformational leadership or effective managerial practices?. *Group & Organization Management*, *23*(3), 220-236.

Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of business*, S251-S278.

Vytal, K., & Hamann, S. (2010). Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. *Journal of cognitive neuroscience*, *22*(12), 2864-2885.

Yukl, G. (2012b). *Leadership in organizations (8th Ed.)*. Harlow, UK: Pearson Education.

Yukl, G. (2013). Leadership in Organizations (8.) Harlow, England: Pearson Education Limited.

Appendix A

MLQ
Transformational Leadership Scale*

Rating scale:

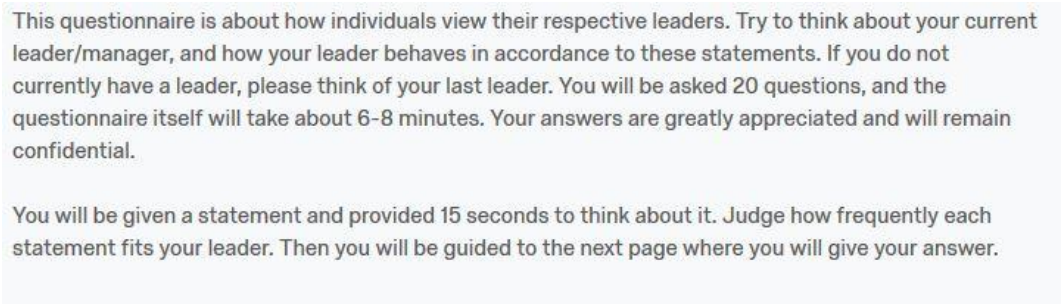
O Not at all **O** Once in a while **O** Sometimes
O Fairly Often **O** Frequently, if not always

1. My supervisor re-examines critical assumptions to question whether they are appropriate
2. My supervisor talks about my most important values and beliefs
3. My supervisor seeks differing perspectives when solving problems
4. My supervisor talks optimistically about the future
5. My supervisor instills pride in others for being associated with me
6. My leader talks enthusiastically about what needs to be accomplished
7. My supervisor specifies the importance of having a strong sense of purpose
8. My supervisor spends time teaching and coaching
9. My supervisor goes beyond self-interest for the good of the group
10. My supervisor treats others as individuals rather than just as members of a group
11. My supervisor acts in ways that build others' respect
12. My supervisor considers the moral and ethical consequences of decisions
13. My supervisor displays a sense of power and confidence
14. My supervisor articulate a compelling vision of the future
15. My supervisor considers an individual as having different needs, abilities, and aspirations from others
16. My supervisor makes others look at problems from many different angles
17. My supervisor helps others develop their strengths
18. My supervisor suggests new ways of looking at how to complete assignments
19. My supervisor emphasizes the importance of having a collective sense of mission
20. My supervisor expresses confidence that goals will be achieved

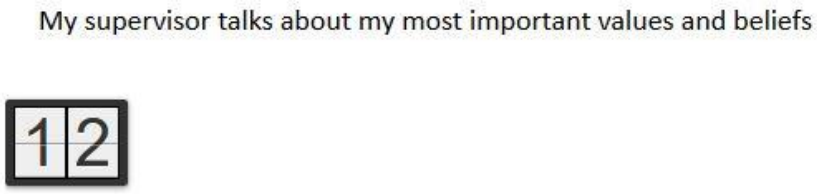
*The numbering on these items were changed from the original MLQ to have better control over the ratings.

Appendix B

Appendix B.1
Introduction to the Survey-- Participants are shown this slide before continuing on to the survey



Appendix B.2
Question Format Example



Appendix B.3
Answer Format Example