Title:

The failing measurement of attitudes: How semantic determinants of individual survey responses come to replace measures of attitude strength

Authors:

Jan Ketil Arnulf. Affiliation: BI Norwegian Business School, Nydalen, N-0442 Oslo, Norway. E-mail: jan.k.arnulf@bi.no, phone +47 46410706

Kai Rune Larsen, Affiliation: Leeds Business School, University of Colorado at Boulder. E-mail: kai.larsen@colorado.edu, phone +1 (720) 938 2436

Øyvind Lund Martinsen, Affiliation: BI Norwegian Business School, Nydalen, N-0442 Oslo, Norway. E-mail: oyvind.martinsen@bi.no, phone: +47 46410737

Thore Egeland. Affiliation: Norwegian University of Life Sciences, 1432 Ås, Norway. E-mail: thore.egeland@gmail.com

Running head: Separating semantic determinants from attitude strength

Word count (excluding figures, tables and references): 9,573

Short author presentations:

Jan Ketil Arnulf, Ph.D., is professor at BI Norwegian Business School, teaching and researching leadership and leadership development.

Kai R. Larsen, Ph.D., is an associate professor of Information Systems at Leeds Business School, University of Colorado at Boulder. His research is aimed at creating a transdisciplinary "backbone" for theoretical behavioral research. He teaches Business Analytics.

Øyvind L. Martinsen, Ph.D., is professor at BI Norwegian Business School in Oslo, Norway. He conducts research in leadership, personality and creativity, and also teaches these issues as well as psychometrics.

Thore Egeland, Ph.D., is professor at the Norwegian University of Life Sciences. He is a statistician and has done pioneering research statistical models for DNA analysis and forensic medicine.

The failing measurement of attitudes: How semantic determinants of individual survey responses come to replace measures of attitude strength

Abstract

The traditional understanding of data from Likert scales is that the quantifications involved result from measures of attitude strength. Applying a recently proposed semantic theory of survey response (STSR), we claim that survey responses tap two different sources: a mixture of attitudes plus the semantic structure of the survey. Exploring the degree to which individual responses are influenced by semantics, we hypothesize that in many cases, information about attitude strength is actually filtered out as noise in the commonly used correlation matrix. We developed a procedure to separate the semantic influence from attitude strength in individual response patterns and compared these to the observed sample correlation matrices and the semantic similarity structures arising from text analysis algorithms, respectively. This was done with four datasets comprising a total of 7,787 subjects and 27,461,502 observed item pair responses. As argued, attitude strength seemed to account for much information about the individual respondents. However, this information did not seem to carry over into the observed sample correlation matrices. These seem to converge around the semantic structures offered by the survey items. This is potentially disturbing for the traditional understanding of what survey data represent. We argue that an enhanced understanding of how cognitive processes are necessary in responses to surveys is now within reach and could offer a valuable path for improvements in the use of survey data.

Keywords: Semantic analysis, surveys, survey response, Semantic Theory of Survey Response (STSR), attitude strength

The failing measurement of attitudes: How semantic determinants of individual survey responses come to replace measures of attitude strength

What do the numbers in survey statistics measure – what do they represent? A growing series of studies show that the statistical properties of survey research are partly or even largely predictable a priori because of their semantic structures (Arnulf & Larsen, 2015; Arnulf, Larsen, Martinsen, & Bong, 2014; Gefen & Larsen, 2017; Larsen & Bong, 2013; Nimon, Shuck, & Zigarmi, 2015). Statistical patterns in responses may therefore express more of the respondents' cognitive operations than their attitudes, as explained in the recently proposed Semantic Theory of Survey Response (STSR, Arnulf et al., 2014; Arnulf, Larsen, Martinsen, & Bong, 2015b).

In his original work on response scales, Likert (1932) was insistent that the object of measurement should be attitude strength. If the emerging statistics are caused by other psychological operations such as the cognitive processing of items, it would cast doubt about their nature as measurements and pose a problem for interpretation (Lamiell, 2013; Slaney & Racine, 2013). On the other hand, the possibility to explore and measure these cognitive operations may open new applications in survey research, possibly improving the predictive validity of such measurements and expanding the use of such instruments.

The purpose of this study is to show how commonly applied statistical techniques actually may frequently filter out its purported subject matter, i.e., attitude strength, leaving only traces of cognitive language parsing operations, contrary to central assumptions in construct validation (APA, 2009; Cronbach & Meehl, 1955; MacKenzie, Podsakoff, & Podsakoff, 2011). This is important for at least two reasons: Primarily, establishing a model of the cognitive framework activated in survey responses can give us more knowledge about how attitudes and mental models are connected. Secondarily, it is principally a serious concern for the interpretation of

psychometric models if the numbers are representing a different phenomenon than commonly purported.

Since Likert published his scales for assessing attitudes 80 years ago (Likert, 1932), they have turned into one of the most prevalent methods of social science in general and organizational behavior (OB) in particular (Cascio, 2012). Respondents assign numbers (or answer in categories later translated into numbers), and the numbers should reflect the degree to which they endorse the survey items (MacKenzie et al., 2011; Michell, 2013; Nunnally & Bernstein, 2010). Increasingly sophisticated statistical procedures are applied to analyze patterns in the ensuing statistics, and common to many of these is that they take the covariance or correlation matrix as their point of departure (Borsboom, 2008; Jöreskog, 1993). Yet, researchers have intermittently published suspicions concerning the role of cognitive processes involved in reading and processing the items. These may influence the observed statistics, distort measures or even create artificial numbers in cases where respondents do not really hold any attitudes (Borsboom, 2009; Feldman & Lynch, 1988; Maul, 2017; Schwarz, 1999).

With the development of digital text analysis in recent years, the cognitive properties of survey items are now open to direct assessment. Recent studies have applied various semantic algorithms to compute overlap in meaning among frequently used Likert scales across fields such as OB and information science (IS). Text algorithms alone were able to explain between 54 and 86% of the variation in correlation matrices from leadership research (Arnulf et al., 2014), displayed comparable factor structures in surveys on IS (Gefen & Larsen, 2017), and demonstrated wide semantic overlap between employee engagement and job satisfaction (Nimon et al., 2015). Semantic properties do not always seem to play an important role, as in the personality test NEO-FFI where semantics had negligible predictive value (Arnulf et al., 2014).

There is yet no theoretical or empirical framework to explain the reason why stable personality traits such as the Big Five may not be as semantically predictable as the responses to surveys on perceptions in organizations.

This study aims to explore and explain why semantics may influence survey statistics by developing a way to track the influence of semantic operations in individual response patterns. By using an individual level analysis, we can track the sources of influence on individual response patterns and determine which aspects of individual respondents that are most likely to accumulate and determine the ensuing sample statistics. The data materials in this study will be the same four datasets as used in Arnulf et al.'s (2014) study, but this time broken down to the level of individual responses. This kind of re-analysis is important because the previous study merely documented the influence of semantics on the sample statistics. To explain why semantics may explain between the negligible and 86% of observed variation in samples, the separate sources of variation in individual responses need to be assessed. For example, if responses measuring stable personality traits are indeed different from responses to shifting perceptions, we need more incisive methods to explore and explain this. The present study is also a methodological innovation, as there is no previously established way to disentangle semantics from attitude strength in survey data. For this reason, we also refrain from stating hypotheses, but stick to outlining the methodological steps and report the ensuing findings.

Attitude Strength and Semantic Relationships

The relationship between semantics and attitude strength is well illustrated by Bertrand Russell's explanation of Wittgenstein's "Tractatus Logico-Philosophicus" (Russell, 1922). Here, the two philosophers make a distinction between three types of scientific facts: Empirical, psychological and logical facts.

Take as an example the sentence "It is raining." The "empirical" fact would be a demonstration of whether it is raining or not. A "psychological" fact would be whether someone believes it to be raining. The "logical" fact, however, is likely to be taken for granted by many people, but still reflects a cognitive process in the brain: The expression "It is raining" needs to be logically comprehensible and recognizably different from other statements. It is, de facto, a prerequisite for the other two, believing in rain and possibly falsifying its empirical justification.

Generally, the object of empirical science is to establish empirical facts such that people with different opinions (in the state of harboring different psychological facts) can subsequently agree on the empirical support for their different beliefs. But in order to determine the nature of their (dis-)agreement, people need to state their beliefs in terms clear enough to display their differing opinions (Wittgenstein, 1922).

This distinction is no mere hair-splitting, it concerns the interpretation of latent variable modelling of survey data (Bagozzi, 2011; Borsboom, 2008). In surveys, people are asked about their attitudes or opinions – "psychological facts" in Russell's terminology. The ensuing data are taken as measures of "empirical facts," used to test social scientific hypotheses about how such empirical facts are related (Shadish, Cook, & Campbell, 2002). If the detected patterns can be shown to be created by semantics (i.e., the mutual relationships of meaning in the items), the data are probably not measuring the purported "empirical" domain, but instead, what is represented are the semantic relationships (or logical relationships in Russell's and Wittgenstein's vocabulary).

In light of modern psychology, we need to make an adjustment to Russell's and Wittgenstein's terminology. Strictly speaking, the "logical" relationships may also be regarded as

workings of psychological mechanisms, but these are of a fundamentally different kind than attitudes. The ability to read and recognize meaning in verbal statements is also a psychological process, but it is created by the language parsing mechanisms in the cerebral cortex (Jackson, Hoffman, Pobric, & Ralph, 2016; LaPointe, 2013; Van Berkum, Koornneef, Otten, & Neuwland, 2007).

The capacity of the cerebral cortex for processing semantic structures in language is central to logical reasoning, but does not in itself process attitudes or truth values of statements. This is necessary for our ability to assess statements with which we do not agree or even products of mere fantasy. The assessment of attitudes and agreement are linked to a broader range of cerebral structures including sub-cortical emotional processing (de Vries, Byrne, & Kehoe, 2015; Fisher, Island, Rich, Marchalik, & Brown, 2015; Honey, Thomson, Lerner, & Hasson, 2012; Nakagawa, Takahashi, Okada, Matsushima, & Matsuda, 2015; Regev, Honey, & Hasson, 2013). For analytical reasons, it is important to keep these two domains apart, and so in the following we will keep the distinction between the three types of facts: Empirical facts established by scientific observation, psychological facts concerning what subjects believe, and logical facts that are simply constituted by coherent and meaningful propositions unrelated to their truth or credibility.

The core element in our method is to use natural language parsing algorithms to compute estimates of how the meaning of survey items overlap. We are using two types of algorithms, one called Latent Semantic Analysis (LSA) (Landauer, 2007) and another termed MI after its inventor (Mihalcea, Corley, & Strapparava, 2006a; Mohler & Mihalcea, 2009). The nature of these will be more closely described in the methods section below. The output of these algorithms is structurally similar to the correlation matrix for the survey variables, but the numbers do not in any way represent attitude strength. Instead, they are simply estimates of the

degree to which the items contain similar propositions. We call the output "semantic similarity indices" and they are similar to what we described as "logical facts" above, in that they only represent the cognitive similarity of statements.

To achieve the aim of this study, we must find a way to differentiate the influence of item semantics from attitude strength in individual survey response patterns. We do this by computing two different response matrices for each respondent – one that is sensitive to response strength, and one that is insensitive to response strength but closer to mere semantic influence.

The rationale behind these two types of matrices starts with how the Pearson product-moment correlation formula picks up the individuals' contributions, as this is the central condition for what happens in further steps of the analysis. The individual contribution in this formula consists of *multiplying* two variables x and y for every participant, thus the name "product-moment," and preserves the magnitude of the variables in each subject. The other elements in the formula simply serve to standardize the variables and centralize the scale around zero. In principle, a matrix of correlations for a sample consists of the average, normalized values of x multiplied by y.

If we multiply each response with all other responses on an individual level, we have for each person what we call an *individual co-product matrix*. It is structurally similar to the sample correlation matrix, but contains no information about other people's responses – nothing about the sample variation. It should hence preserve information about the respondents' attitude strength. For example, a person scoring 4 and 5 on x and y will obtain a co-product of 20, but the co-product for a person scoring 1 and 2 will only be 2.

We can compute another, but similar matrix for each person that is relatively de-coupled from the individual's attitude strength. By *subtracting* the score on each variable from every other variable and keeping the absolute difference, we obtain what we call an *individual item-distance matrix*. Again, it is structurally similar to the sample correlation matrix, displaying a numerical value for each combination of variables. This matrix contains little or no information about the subject's attitude strength. To repeat the example above, the item distances for a person scoring 4 and 5 will be 1, but the distance score for a person scoring 2 and 1 will also be 1.

The *individual item distance matrix* is theoretically akin to the semantic similarity matrix because of two conditions regarding the distances: a) Any two items containing the same proposition should logically be rated in the same way regardless of the subjects' endorsements of the items. B) Conversely, for two items to have different scores, the respondents would need to perceive them as different. So, while there may be other reasons why items are scored similarly, these two conditions together should make the item-distance matrix converge around how *similar* any subject perceives these items, regardless of the subject's attitude strength. In contrast, the same individual's *co-product matrix* would express similar information but with the added information of the individual's attitude strength.

The purpose of these two matrices is to compare how similar the responses of each individual are to our two criteria: the observed sample correlation matrix and the semantic similarity matrix. We may then proceed to explore the relative impact of semantics and attitude strength for each subject, and for each sample in our study.

As it may be hard for the reader to keep track of the different types of matrices, we have created a sample for comparison in Table 1. It displays the statistics for a survey scale with six

items measuring intrinsic motivation, originally published by Kuvaas (2006b). The wording of these items are: 1. The tasks that I do at work are themselves representing a driving power in my job, 2. The tasks that I do at work are enjoyable, 3. My job is meaningful, 4. My job is very exciting, 5. My job is so interesting that it is a motivation in itself, 6. Sometimes I become so inspired by my job that I almost forget everything else around me

In Table 1, the correlations among these items in the sample from our 2nd dataset (N=255) are displayed on the top left. The matrix on the top right displays the semantic similarity indices (MI values) for the same items. On the bottom, we have displayed the item-distance matrix (left) and the co-product matrix (right) for a randomly selected respondent in this sample who scored 5,5,4,5,5 and 4 on these six items:

Table 1

Examples of the four different matrices used in this study.

Table 1: Examples of the four different matries used in the study

Sample correlations matrix

	Item #1	Item #2	Item #3	Item #4	Item #5
Item #2	.67				
Item #3	.52	.58			
Item #4	.59	.71	.53		
Item #5	.68	.76	.53	.79	
Item #6	.46	.52	.36	.49	.60

Semantic MI-values matrix

	Item #1	Item #2	Item #3	Item #4	Item #5
Item #2	.78				
Item #3	.73	.73			
Item #4	.64	.73	.83		
Item #5	.59	.61	.81	.66	
Item #6	.52	.56	.63	.56	.55

Distance matrix for Respondent 002

	Item #1	Item #2	Item #3	Item #4	Item #5
Item #2	0				
Item #3	1	1			
Item #4	0	0	1		
Item #5	0	0	1	0	
Item #6	1	1	0	1	1

Co-product matrix for Respondent 002

	Item #1	Item #2	Item #3	Item #4	Item #5
Item #2	25				
Item #3	20	20			
Item #4	25	25	20		
Item #5	25	25	20	25	
Item #6	20	20	16	20	20

The two different computations may seem a bit confusing at first, so a quick recapitulation and comparison may be helpful. The main purpose is to explore how individual response patterns are informed by semantics, and how this influence contributes to the sample statistics. The co-product procedure of multiplying all items with all other items was chosen for two main reasons: Primarily because we will use the sample correlation matrix as a criterion for our further analyses (this, or its closely related derivative the co-variance matrix, being the point of departure for much psychometric modelling). The co-product is the individual value that comes closest to the individual's contribution to the sample correlation matrix. Hence, the co-product should ideally be the values informing the correlation matrix. Secondly, there is a frequent assumption in quantitative models that factors interact to create their effects. As an example, the "interaction term" in multiple regression is usually entered as a multiplication of the interaction variables (Aiken, West, & Reno, 1991). Based on these arguments and traditional practice, the co-products are the most obvious candidate for examining the individual's contribution to sample data patterns based on measurements of attitude strength.

On the contrary, the item differences that come from subtracting the item scores from each other are practically devoid of information about attitude strength. While one may object that difference between the scores 4 and 3 depend on attitude strength, the difference is the same as between 4 and 5 or 2 and 1. From the point of view of attitude measurement, the distances take the "quanta" out of the numbers, as can be seen in Table 1.

The strength of attitudes is important because it could conceivably signal a propensity for action if the attitude strength reaches a certain level, which has been a core problem in attitude research since the 1930s (Fishbein & Ajzen, 1975; Kim, Hahn, & Lee, 2015; LaPiere, 1934; Likert, 1932; Prinz, Aschersleben, & Koch, 2009; Triandis, 1980; Zimbardo, Ebbesen, &

Maslach, 1977). If the recordings of attitude strength are only used to compare similarity of responses, the information about attitude strength is left out of the picture. The item distances represent information about the similarity between responses without considering the absolute level of these responses. Seen as item distances, the Likert scale numbers simply provide a medium for expressing proximity in meaning.

A main proposition of this study is that we can separate attitude strength from semantic processing in individual records by computing these two matrices for each respondent. We can explore if the compliance with semantic structures is related to attitude strength, or if this information is simply a product of response proximities. We can do this by exploring their relationships to the two criteria: the semantic similarity matrix which we know is unrelated to attitudes (because it is created by a computer using the text alone), and the sample correlation matrix which is a result of respondents expressing their attitudes.

The sample correlation matrix commonly serves as the input for further modeling, together with its derived relative the covariance-matrix (Abdi, 2003; Fornell & Larcker, 1981; Jöreskog, 1993; Jöreskog & Sörbom, 1993; Marsh et al., 2010; P.M. Podsakoff, MacKenzie, & Podsakoff, 2012b). To the extent that this data structure contains information about attitude strength, it should be statistically more strongly related to the individual subjects' co-product matrices. Conversely, the item distance matrices should be more strongly related to the semantic values matrix. The relative influence of semantics on each subject would then be the degree to which the response patterns of individuals are more closely resembling the semantic patterns than the sample correlation matrix.

The sample correlation matrix and the semantic values matrix are used here and referred to as "criteria" because of the following conditions: If a single respondent's matrix is structurally similar to the sample correlation matrix, then this person is simply responding in a statistically average way – the pattern that emerges is characteristic of the whole sample. With varying departures from this pattern, the respondent is responding in a way that differs from the statistically common pattern. Conversely, if a person's matrix is similar to the semantic values matrix, he or she is in accordance with what is semantically expected. We can therefore explore at least three characteristics of individual responses: a) The degree to which their responses are congruent with what is semantically given, 2) the degree to which patterns in their attitude strength is similar to what is semantically given, and 3) which aspect of their response pattern – attitude strength or semantic congruence – is more likely to be contributing to the final statistics for the whole sample.

The various matrices can be compared as relationships visualized in Figure 1. The sample correlation matrix (upper left triangle) includes the empirically obtained numbers from the whole sample. The semantic matrix (upper right triangle) has the same structure of items, but their relationships are only computed by semantic algorithms and no responses are involved. The two individual matrices (lower central triangles) exemplify the same structures for individuals, obtained once as a co-product matrix (bottom left) and once as an item-distance matrix.

Figure 1: The four types of data matrices used in this study and their mutual relationships
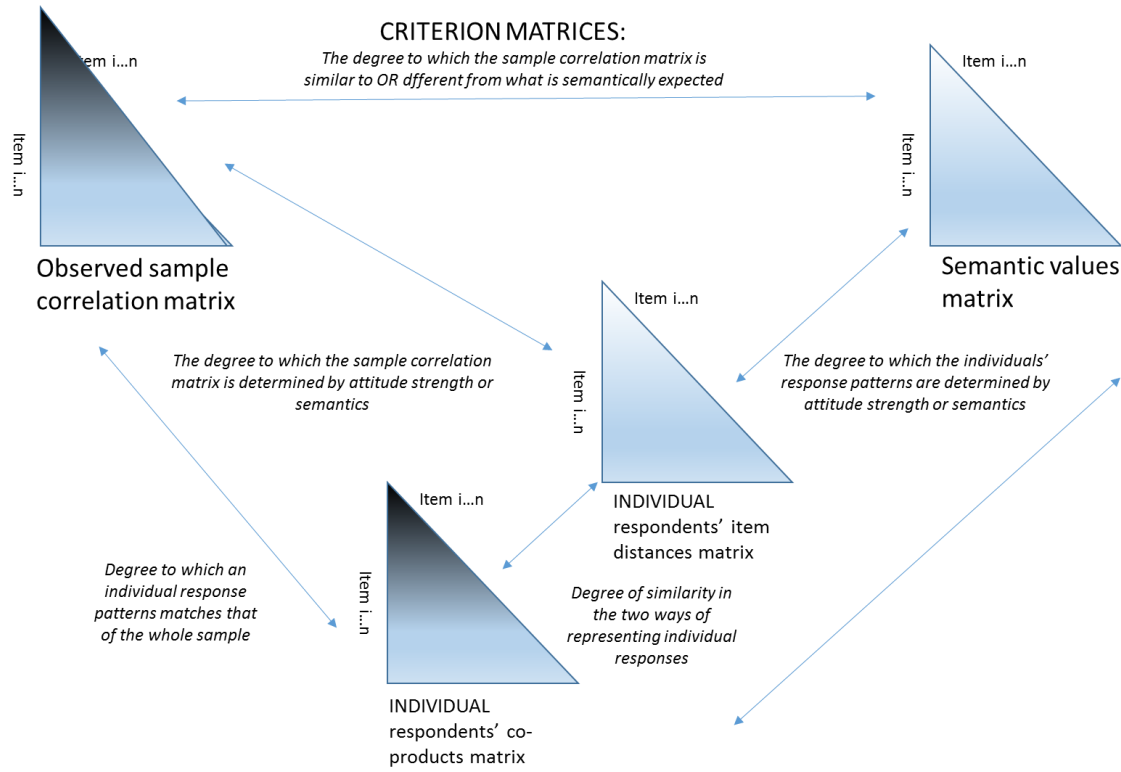
STSR and Linguistic Algorithms

We have assessed the semantic properties of the items using several available algorithms, primarily focused around LSA and WordNet-based approaches (e.g., Dennis, Landauer, Kintsch, & Quesada, 2003; Landauer, 2007; Larsen & Monarchi, 2004; Larsen, Nevo, & Rich, 2008; Li, McLean, Bandar, O'Shea, & Crockett, 2006; Mihalcea et al., 2006a). They share the commonality of only analyzing pure text (i.e., no data about how respondents may endorse propositions in these texts).

Text algorithms are a complex topic and a detailed presentation of the procedures involved is not possible within the present format, so we will only describe them on a superficial

level below. For more details, we must refer to the appendix of Arnulf et al. (2014) or more technical sources such as Dennis et al. (2013) for LSA or Mihalcea, Corley, and Strapparava (2006b). A quick and user-friendly introduction can be found at lsa.colorado.edu for LSA. An approach more suitable for analyses like the present study can be found at http://sims.theorizeit.org/.

When comparing individual responses to semantics, we use only a single text algorithm termed MI, originally developed by Mihalcea et al. (2006b; Mohler & Mihalcea, 2009). We do this to avoid complexities in establishing a joint semantic matrix with values from several types of algorithms. From the previous studies, the MI values seem to be the single most powerful predictor of the responses (Arnulf & Larsen, 2015; Arnulf et al., 2014). MI is particularly well-suited for short sentences and derives its information about the meaning of words from a semantic knowledge base and term ontology called WordNet (Miller, 1995; Poli, Healy, & Kameas, 2010). In WordNet, words are organized into a hierarchical structure (an upside-down tree) with synsets (groups) of synonymous words throughout the tree. The further down in the tree a word occurs, the more specific it is, so two words that are part of the same "branch" but at different depths have different specificity (e.g., pear->edible fruit->produce). Words may be at the same depth in the hierarchy but in different branches. For example, the words car and pear may be similar in that they are as far down in their branches as WordNet will go (Asian Pears and BMW models are not included). WordNet allows word similarity to be computed by measuring the path length and path depth between two words in the database. MI's word specificity juxtaposes subtle nuances of meaning (e.g., apple and pear) and generic conceptual meaning (e.g., fruit and vegetable). WordNet is a general lexical database, and MI therefore reflects neutral, "everyday" language and lacks domain-specific knowledge.

The MI algorithm first identifies single words in each sentence and computes similarity as the shortest distance between these words' synsets (sets of synonymous words) in WordNet (Mihalcea et al., 2006a; Mohler & Mihalcea, 2009). Word-similarity scores are taken as inputs for a formula on sentence-level similarity. Thus, MI is built first on word-level meaning, moving up to sentence-level structures embedding the words. The output of MI is a similarity score between 0 and 1, where 0 indicates no semantic overlap and 1 indicates an exact match. In the following, these numbers will be referred to as "semantic similarity indices."

Semantic similarity indices are then computed through the MI algorithm as the semantic distance between any pair of items (Item1-Item2), (Item1-Item3), etc. The resulting output is a matrix representing the semantic relationships across all items. The following two items may serve as examples: "*Generally, I like to work here,*" and "*I would be very happy to spend the rest of my career with this organization.*" In our surveyed data, these are found to have a correlation of .51, and their MI similarity index is .50 (a magnitude equivalent to their correlation).[1]

However, when estimating the influence of semantics on the whole sample, we can use multiple regression to complete the MI values with the values from LSA used in previous analyses (Arnulf et al., 2014). LSA is different from MI in that it "demonstrates a computational method by which a major component of language learning and use can be achieved" (Landauer, 2007, p. 31). The way LSA extracts "meaning" from sentences is by relying on huge volumes of text fragments organized along a specified number of dimensions (a usual number of dimensions is 300). The multi-dimensionally represented texts are called "semantics spaces" in which all terms have specific locations, represented as vectors. New combinations of words can be mapped

---

[1] It should be noted that we do not generally expect MI similarity indices to be interchangeable with correlations, but that insofar as semantics have influenced the correlation matrix, there should be a measurable relationship between the two.

as vectors in these spaces, resulting in quantitative measures of their similarity with other expressions. For this reason, LSA has by some of its pioneers been argued as a mathematically based theory of meaning (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Landauer, 2007). It is possible for LSA to determine that two sentences share the same meaning even if they do not share words or appear in the same passage. The reason for this is that the aggregated patterns of the semantic spaces will indicate the intended meaning of a sentence relative to all other possibilities in the semantic space.

Comparing two texts "Doctors operate on patients" with "physicians do surgery," LSA returns a cosine of .80. Comparing "The radius of spheres" with "a circle's diameter" the cosine returns .55, but between "The radius of spheres" and "the music of spheres," the cosine is as weak as .01 (Dennis et al., 2003). Because these methods have been presented in detail elsewhere (Arnulf et al., 2014; Gefen & Larsen, 2017; Larsen & Bong, 2016; Nimon et al., 2015), we refrain from elaborating on LSA here. The main point is that each of these algorithms have limitations so that a bundle of MI values and LSA extracted cosines performs better in predicting human language parsing than each of them in isolation.

Attitude Strength or Semantics – Does it Matter?

At first glance, one may ask whether the difference between semantics and attitude strength matters in practice, here operationalized as item distances and co-products. Many researchers are used to constructing measurement scales with different wordings of the same items. This is a frequent way to ensure that the scale taps different facets of a construct while keeping it internally consistent and delimited from other measures included in the instrument (Nunnally & Bernstein, 2010). So far, the semantic similarities would be expected.

The real problem however appears if the other constructs tested are similarly determined by the semantic relationships, such that there is a necessary semantic relationship between independent and dependent variables. In that case the "factors" are not separate, but still only facets of the same phenomenon. To prevent this, researchers commonly apply various techniques such as factor analysis to ensure the relative independence of these measures (Abdi, 2003).

The theoretical insufficiency of this assumption was poignantly outlined by van Knippenber and Sitkin (Van Knippenberg & Sitkin, 2013), who showed that the construct of "transformational leadership" conflates the independent with the dependent variable through definitions and operationalizations. Empirically, the tautological link between leadership and other constructs was demonstrated in the first paper on STSR (Arnulf et al., 2014), where the relationship between all constructs was found to be determined by semantics notwithstanding the rotated factor structures.

If the research models simply end up with "what can be said," this may explain why there are often inflated statistics and a lack of predictive validity from such models (Bagozzi, 2011; Gollwitzer & Sheeran, 2006; P.M. Podsakoff, MacKenzie, & Podsakoff, 2012a; P. M. Podsakoff & Organ, 1986). From a research point of view, it may actually be more interesting to see if attitude strength may override cognitive restraints on behavior. Theories on organizational behavior frequently hypothesize that affective responses to leaders, job characteristics and social environments reduce calculative behaviors (e.g., Conger & Kanungo, 1998; Felfe & Schyns, 2014; Ilies, Nahrgang, & Morgeson, 2007; Kuvaas, Arnulf, & Dysvik, 2008). For research to make progress in this field, it is important that methods distinguish between the cognitive, semantic relationships between words describing workplaces and the affective or attitudinal dynamics that shape loyalty, commitment or their negative counterparts.

The Problem of Signs

MI values do not take negative values at all, and when LSA cosines are negative, it does not signal a contrary statement. Thus, the algorithms cannot detect that an item is reversed, as is sometimes the case in surveys. Some survey structures depend heavily on negative relationships and so finding a way to deal with this is important to estimate the actual relationship between the semantic and the surveyed matrices.

In this study, we follow the same procedure as described previously by Arnulf et al. (2014), simply correcting the sign of the semantic values for reverse-scored items. One special case is the Multifactor Leadership Questionnaire (Avolio, Bass, & Jung, 1995) which does not contain reversed items. Instead, about one fourth of the 990 item pairs were negatively correlated because their subscales are theoretically devised to be the opposite of effective leadership ('laissez-faire' and 'passive management by exception'). As described elsewhere (Arnulf et al., 2014), their negative relationships can be argued a priori, similarly to using one-tailed instead of two-tailed significance tests. We tested this argument by making signs of semantic identity negative when items combine positive leadership with the subscales 'laissez-faire' and 'passive management by exception' and thereby correctly identified 255 of the 264 negative correlations ($p < .001$).

Samples and Measures

We compared individual response patterns from the following four datasets, establishing a sample correlation matrix and a semantic similarity matrix for each dataset. For each respondent in each sample, we computed their respective co-product and item distance matrices.

Dataset 1

A sample of 1,220 respondents filled out the Multifactor Leadership Questionnaire (MLQ) (Avolio et al., 1995) rating their immediate superior manager. The original sample consisted of 1,649 persons from a Norwegian bank who responded to a Norwegian translation of the survey. In the present analysis, we only included respondents for whom the demographics were known. Overall, 48.9% were females, the sample age mean was 46 years, 71% were not managers, middle managers made up 26% and 2% listed themselves as top managers. The MLQ contains 45 items, yielding a matrix of 990 unique pairs of items.

Dataset 2

This contains responses from 255 scientists and engineers in a governmental research agency. They responded to the MLQ (Avolio et al., 1995) rating their own superior managers, but in addition responded to scales measuring perceptions of economic and social exchange (7 and 8 items each) (Shore, Tetrick, Lynch, & Barksdale, 2006), six items measuring intrinsic motivation (Kuvaas, 2006b), seven items measuring organizational citizenship behavior (OCB) (Van Dyne & LePine, 1998), five items measuring turnover intention (TI) (Kuvaas, 2006a), as well as ten items measuring work quality and work effort (Kuvaas, 2006b). In total, the survey contained 79 items producing 3,081 unique pairs in each of the matrices. Demographics revealed 66.7% were male with a mean age of 38 years.

Dataset 3

In this sample, 981 employees from the Norwegian Armed Forces responded to a broad range of diverse items concerning leadership and other OB variables rating their own superior managers: 20 items from the MLQ (Avolio et al., 1995); eight items measuring leader-member-exchange (LMX) (Bernerth, Armenakis, Feild, Giles, & Walker, 2007); 20 items from the Leadership Behavior Development Questionnaire (LBDQ) (Stogdill, 1963) where 10 items are related to initiation of structure and 10 items to consideration. Additionally, eight items were from a scale measuring affective organizational commitment (Meyer, Allen, & Smith, 1993), job satisfaction (three items) (Cammann, Fichman, Jenkins, & Klesh, 1979), TI (Kuvaas, 2006a), and a scale measuring self-rated work quality and work effort with altogether 10 items (Kuvaas, 2006b). In total, there were 69 items in this survey providing 2,346 unique pairs of items in the various matrices computed. No demographics were available in this sample.

Dataset 4

This consisted of a sample of 5,332 students responding to an officially translated Norwegian version the NEO-FFI, which is a commonly used five-factor personality inventory (Costa & McCrae, 1992; Martinsen, Nordvik, & Østbø, 2003). This is a short form with 60 items, yielding 1,770 unique pairs of correlations in the various matrices. Demographics for the sample showed the mean age to be 25 years, and 44.7% were male.

Analytic Procedures

We computed the co-products and distances matrices for all respondents in all four datasets. This included a total of 7,787 respondents supplying one co-product matrix and one distance matrix each, yielding a total of 27,461,502 observed item pair responses. For each respondent, we explored how similar each of the matrices were to the two criterion matrices, the

observed sample correlation matrix and the semantic matrix, respectively. To describe similarity, we simply used the Pearson correlation between the individual's matrix and the two criteria. If a person's co-product matrix correlates highly with the correlation matrix (regardless of response strength), this person contributes to the observed pattern in sample statistics (in the form of alphas and factor structures derived from these). On the contrary, if the person's co-product matrix correlates weakly with the observed sample matrix, this person's contribution to the sample contains noise that is filtered out in the sample matrix.

We measured the individual's similarity with the semantic indices in the same way. A person whose individual matrices correlate highly with the semantic similarity matrix can be said to comply with the expected semantic patterns of the survey. In the case of weak correlations, the individual can be said to depart from the semantically expected pattern. As argued above, we expect the semantic similarity index matrix to be closer to the individuals' item distance matrices than to their item co-product matrices.

For each individual, we computed these similarities (correlations between their individual matrices and the two criteria) and retained these for further explorations. We also averaged the co-products and differences across individuals and obtained another two measures: The average item-distance matrices and the average co-products for each of the four samples.

The literature on partialling independent variables sometimes suggests that such comparisons should be done using intra-class-coefficients (Lynam, Hoyle, & Newman, 2006). However, in this case we chose to stay with the simplest procedures possible. The sources we compare are identical constructs (the contents of the scales) and the only differences that we compare are between the calculations of co-products and distances. We are simply exploring

which aspects of the individual response patterns are most similar to the observed sample statistics. In no case did we try to nest analyses as these procedures have not been tried before, and we want to reduce complexities until a solid methodology is established in the field of survey semantics.

Results

Table 2 shows the distributions for co-product and distance similarities. While the values vary, the relationships are strong for the first three datasets containing OB surveys, but break down in the case of the NEO. The differences among all the means are statistically significant with the following exceptions: In samples 1 and 2, the co-product matrices' similarity with the sample matrix are non-significantly different from the distance matrices' similarity with the semantics. In samples 3 and 4, there is a non-significant difference between the co-product and difference matrices in their similarity with semantics.

Generally, there does not seem to be much difference between the distance and the co-products in terms of their distributions. The distance matrices seem to reach higher maximum similarities with the observed correlation matrix as well as with the semantic values, and their minimum values are also not as low. This latter phenomenon may be due to the fact that an individual's co-product matrices can be negatively related to the sample correlation matrix.

Table 2

Descriptives for individual co-product and distance matrices all 4 datasets.

 DATASET 1   *(MLQ only)*

	Co-products similarity with sample correlation matrix	Co-products similarity with semantic matrix	Item distances similarity with sample correlation matrix	Item distances similarity with semantic matrix
Mean	.44	.42	.45	.42
Maximum	.86	.91	.93	.96
Minimum	-.49	-.51	-.17	-.16

DATASET 2   *(MLQ with measures of motivation and outcomes)*

	Co-products similarity with sample correlation matrix	Co-products similarity with semantic matrix	Item distances similarity with sample correlation matrix	Item distances similarity with semantic matrix
Mean	.25	.31	.27	.26
Maximum	.55	.69	.66	.77
Minimum	-.15	-.13	.00	-.06

DATASET 3   *INDIVIDUAL LEVEL (Thre types of leadership, div. motivation, outcomes)*

	Co-products similarity with sample correlation matrix	Co-products similarity with semantic matrix	Item distances similarity with sample correlation matrix	Item distances similarity with semantic matrix
Mean	.23	.35	.37	.34
Maximum	.75	.92	.78	.97
Minimum	-.51	-.56	-.02	-.16

DATASET 4   *NEO-FFI 5-factor personality measures only*

	Co-products similarity with sample correlation matrix	Co-products similarity with semantic matrix	Item distances similarity with sample correlation matrix	Item distances similarity with semantic matrix
Mean	.21	-.01	.23	.02
Maximum	.45	.31	.65	.22
Minimum	-.06	-.33	-.05	-.19

Figures 2 to 5 show the distribution of two individual level variables: How similar the individual co-products are to the correlation matrix (sorted in this order), and the same individual's item distance matrix similarity with semantics:

Figure 2. Dataset 1 MLQ values, sorted by co-product similarity (correlation) with the sample matrix (thick line) and plotted with same individual's distance matrix similarity with semantic (N=1220).
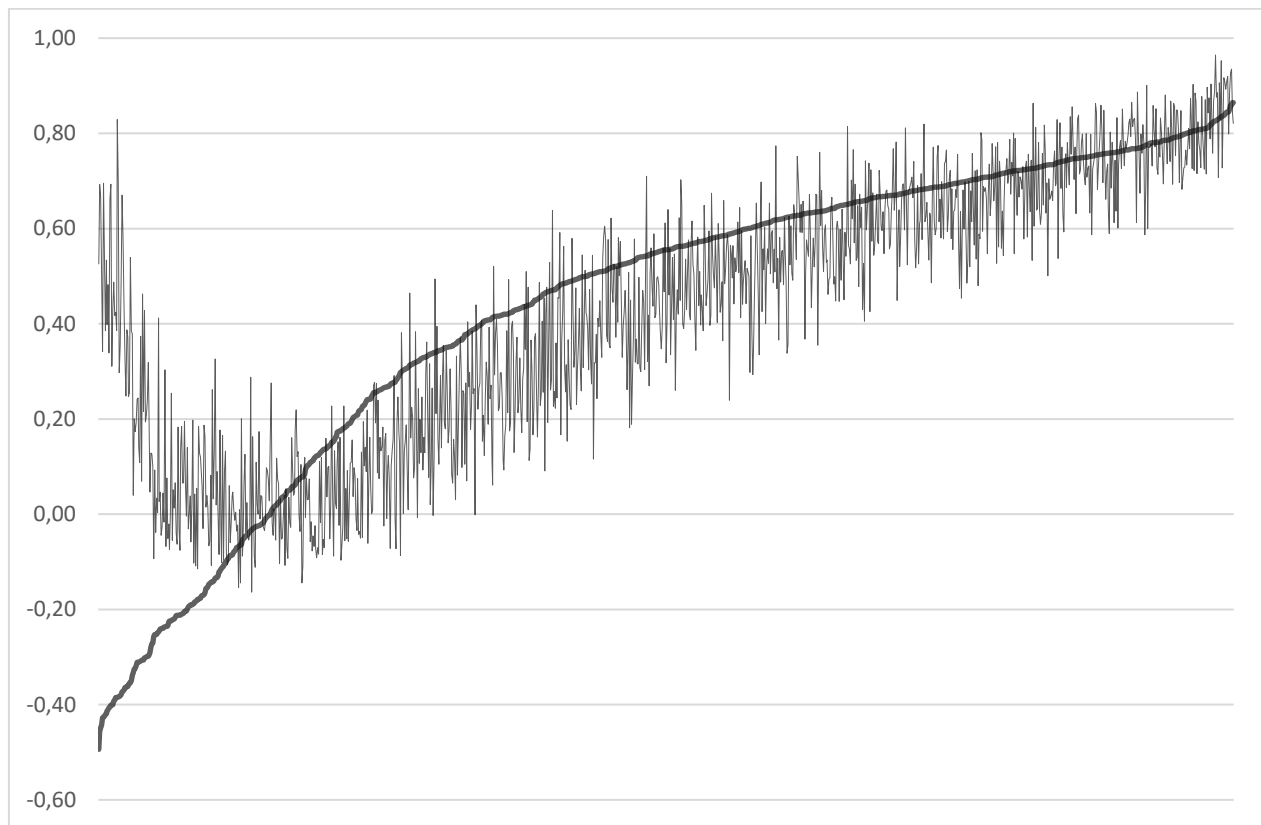
Figure 3. Dataset 2, individual responses MLQ with motives and outcomes, sorted on similarity with sample matrix (thick line) and item distances similarity with semantics (thin line) (N=255).
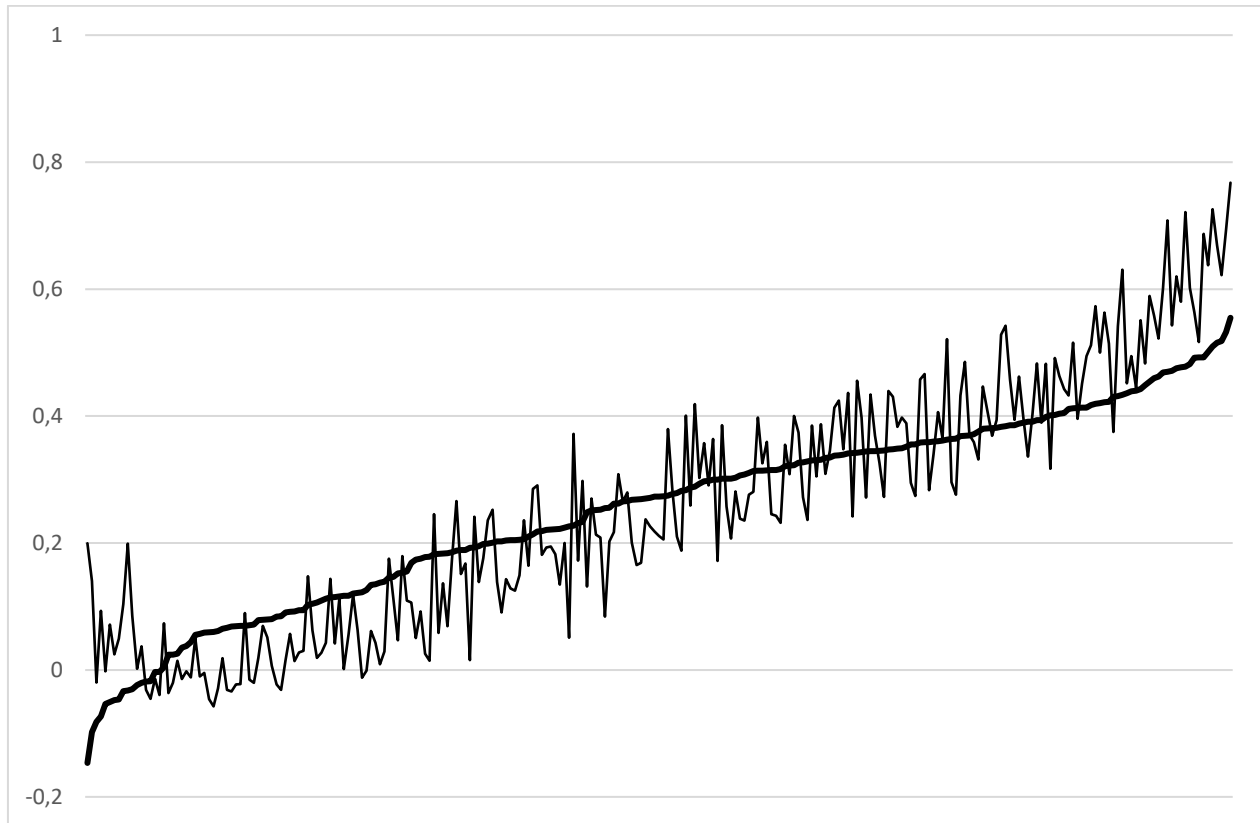
Figure 4. Dataset 3, Individual responses to 3 leadership styles, motives, outcomes, sorted by co-product similarity to sample matrix (thick line) and with distance similarity to semantics (thin line) (N=981)
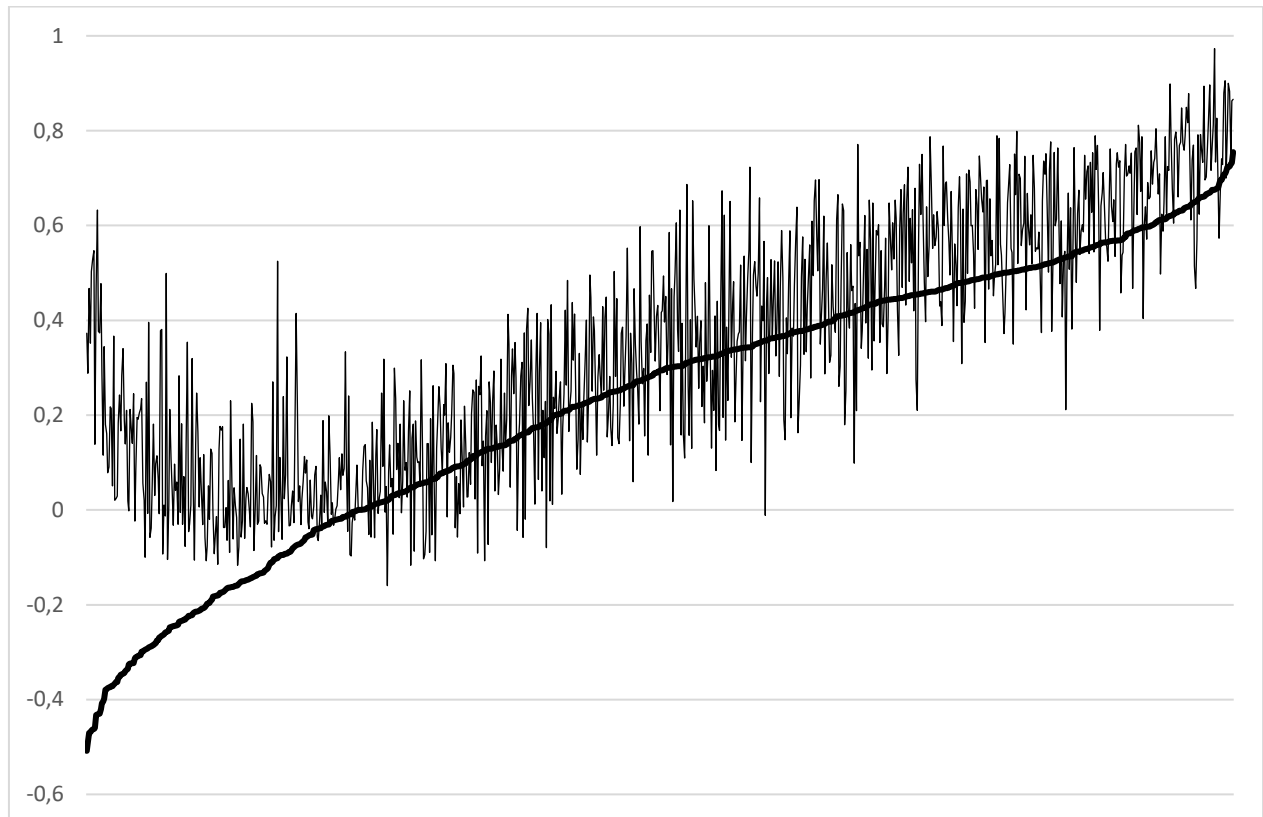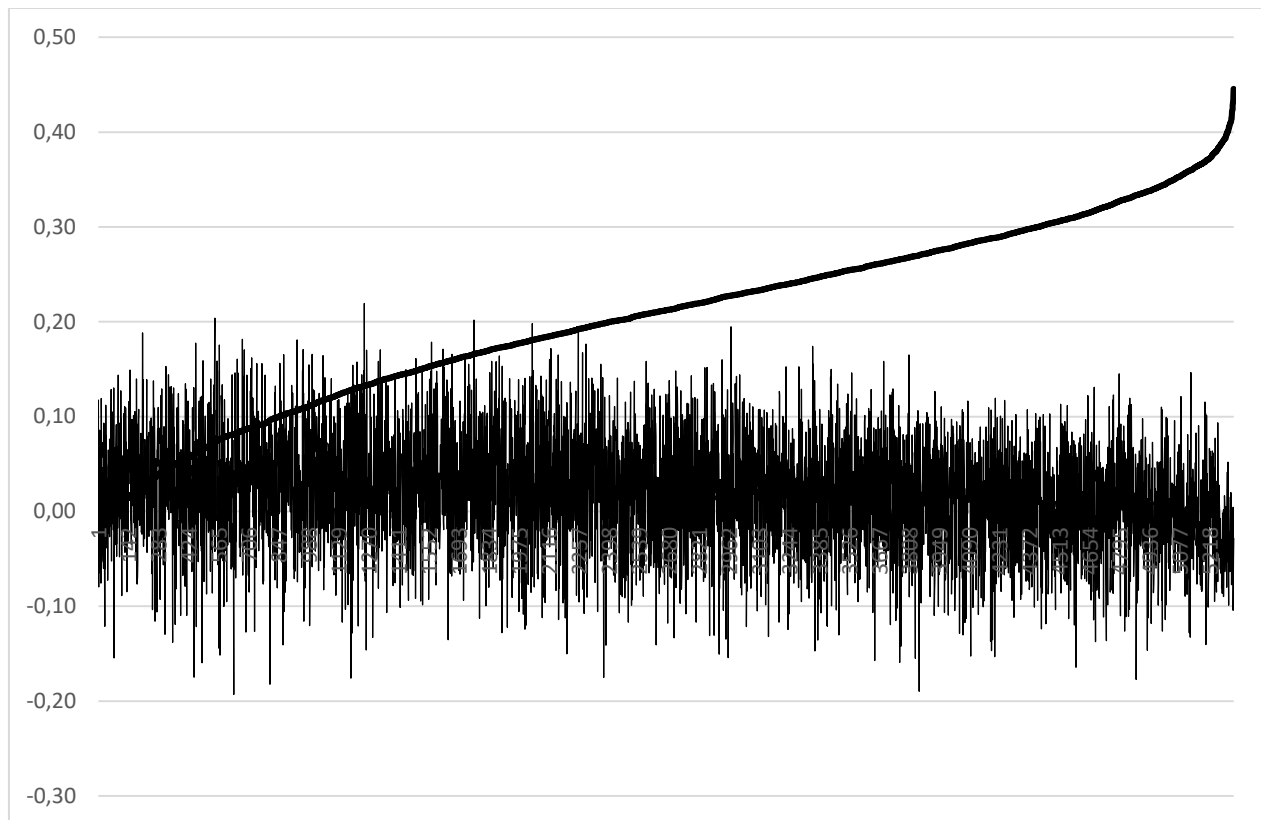
Figure 5. Dataset 4, Individual NEO profiles sorted by similarity with sample correlation matrix (thick continuous line) and item distance similarity with semantics (thin jagged line) (N=5331).

The plots show that for the first three datasets, to the extent that the individual's co-products are similar to the observed correlation matrix, the same person's item distance matrix will be similar to the semantic values. In the fourth dataset containing the NEO data, this phenomenon is not

obviously present. Also, in the first three datasets, an extreme negative relationship between the

co-product matrix and the sample correlation matrix implies higher compliance with what is

semantically expected. This suggests that in the individual responses, there is a relationship

between response strength and the sample correlation matrix.

We therefore turn to an exploration of how the individuals' score levels (indicating

attitude strength) relate to how similar the individuals' co-product and distance matrices are to

the two criteria: sample correlations and semantics. Table 3 displays this for all four datasets.

Table 3 *How relationships between subscale scores (attitude strength) relate to individual co-product and distance matrix properties*

DATASET 1: How co-products and distances related to respondents' scale levels (N=1220, p all values < .01)

Survey subscale	Co-products resembling sample matrix	Co-products resembling MI (semantics)	Distances resembling sample correlations	Distances resembling MI values
Transformational leadership	.86	.83	.70	.67
Transactional leadership	.25	.31	.26	.32
LaissezFaire	-.83	-.85	-.66	-.66

DATASET 2: How co-products and distances related to respondents' scale levels (N=255, p all values < .01)

Survey subscale	Co-products resembling sample matrix	Co-products resembling MI (semantics)	Distances resembling sample correlations	Distances resembling MI values
Transformational leadership	.81	.77	.73	.82
Transactional leadership	.28	.29	.23	.33
LaissezFaire	-.66	-.62	-.63	-.66
Economic exchange	-.60	-.62	-.47	-.44
Social exchange	.41	.45	.35	.36
Intrinsic motivation	.58	.59	.47	.42
Work effort	.36	.38	.32	.30
Work quality	.22	.24	.19	.19
Turnover intentions	-.54	-.59	-.40	-.45
Organizational citizenship behavior	.32	.33	.30	.27
AVERAGE ABSOLUTE VALUES	*.48*	*.49*	*.41*	*.42*

DATASET 3: How co-products and distances related to respondents' scale levels (N=981, p all values < .01)

Survey subscale	Co-products resembling sample matrix	Co-products resembling MI (semantics)	Distances resembling sample correlations	Distances resembling MI values
Initating structure	.52	.44	.35	.45
Consideration	.74	.55	.41	.51
Transformational leadership	.79	.57	.50	.58
Leader-member-exchange	.72	.54	.45	.53
Job satisfaction	.62	.67	.38	.44
Work effort	.19	.29	.22	.18
Work quality	.12	.22	.20	.17
Affective commitment	.53	.67	.47	.54
Turnover intention	-.78	-.92	-.61	-.67
AVERAGE ABSOLUTE VALUES	*.56*	*.54*	*.40*	*.45*

DATASET 4: How co-products and distances related to respondents' scale levels (N=5331, p all values < .01 except marked [ns])

Survey subscale	Co-products resembling sample matrix	Co-products resembling MI (semantics)	Distances resembling sample correlations	Distances resembling MI values
Neuroticism	-.83	.33	-.77	.20
Extraversion	.57	-.16	.52	-.17
Openness to experience	.15	-.45	.09	-.28
Agreeableness	.17	-.11	.16	-.06
Conscientiousness	.64	.01[ns]	.57	.09
AVERAGE ABSOLUTE VALUES	*.47*	*.21*	*.42*	*.16*

The findings in Table 3 show that attitude strength (score levels) are more strongly related to the co-products than to the item distance matrices. This is expected theoretically, both because measuring attitude strength is the purpose of the surveys and because of the correlation coefficient formula. However, for the first three datasets, the scale levels are not only related to how co-products are similar to the correlation matrix – they are also related to how the co-products relate to the semantics. It is as if people who respond strongly to the theoretically favorable conditions in the OB surveys are also more compliant with the semantic patterns, possibly by paying more attention to the exact wording of the survey. Conversely, people expressing more moderate attitude strengths are responding with less compliance to the semantically expected patterns.

This relationship is much weaker for the NEO data, and Table 3 also holds more differential information about possible explanations. It turns out that the personality scores are differently sensitive to semantic values. Table 4 shows that the five traits have a very different impact on an individual's response patterns – both in the way they relate to the sample matrix and to what extent they are influenced by semantics. Not surprisingly, people scoring high on conscientiousness are compliant with semantics, as are people scoring low on openness. These individuals may be more inclined to interpret text in literal terms, hence more semantically predictable. People who score low on neuroticism display a strong tendency to respond very differently from all other people – their responses are negatively predicted by the sample correlation matrix.

Table 4
Each personality trait regressed on respondent's similarity with sample matrix and semantics matrix

	Neuroticism		Extraversion		Openness		Agreeableness		Conscientiousness	
Adjusted R^2	.69		.32		.20		.03		.46	
	St. beta	Sig.	St. beta	Sig.	St. beta	Sig.	St. beta	Sig.	St. beta	Sig.
Co-products similarity with sample matrix	-.81	.00	.58	.00	.00	.80	.15	.00	.71	.00
Co-products similarity with semantics	.07	.00	.03	.03	-.45	.00	-.06	.00	.24	.00

One reason why Neuroticism correlates so highly with semantics in Table 4 may be a noise phenomenon. If many of these individuals respond differently from others, they probably also respond differently from each other. It is possible that the only thing such people have in common is the semantic properties of these items, thus strengthening the relationship between semantic compliance and neuroticism. This is illustrated as a plot in Figures 6 and 7:

Figure 6. NEO, individual respondents' similarity with the sample correlations matrix (thick straight line), plotted with their z-scores on Neuroticism (thin jagged line). (N = 5,331).
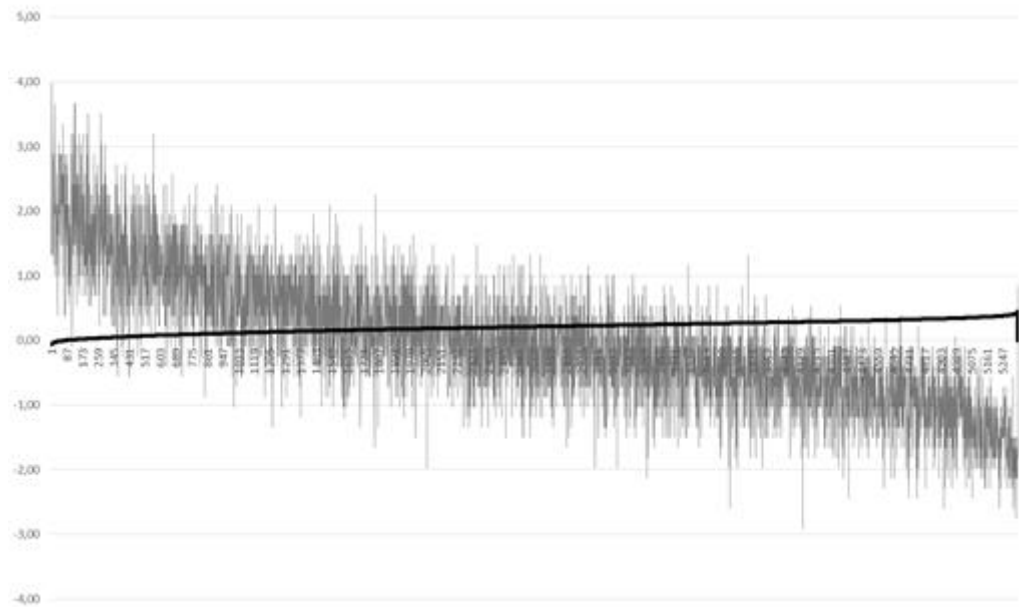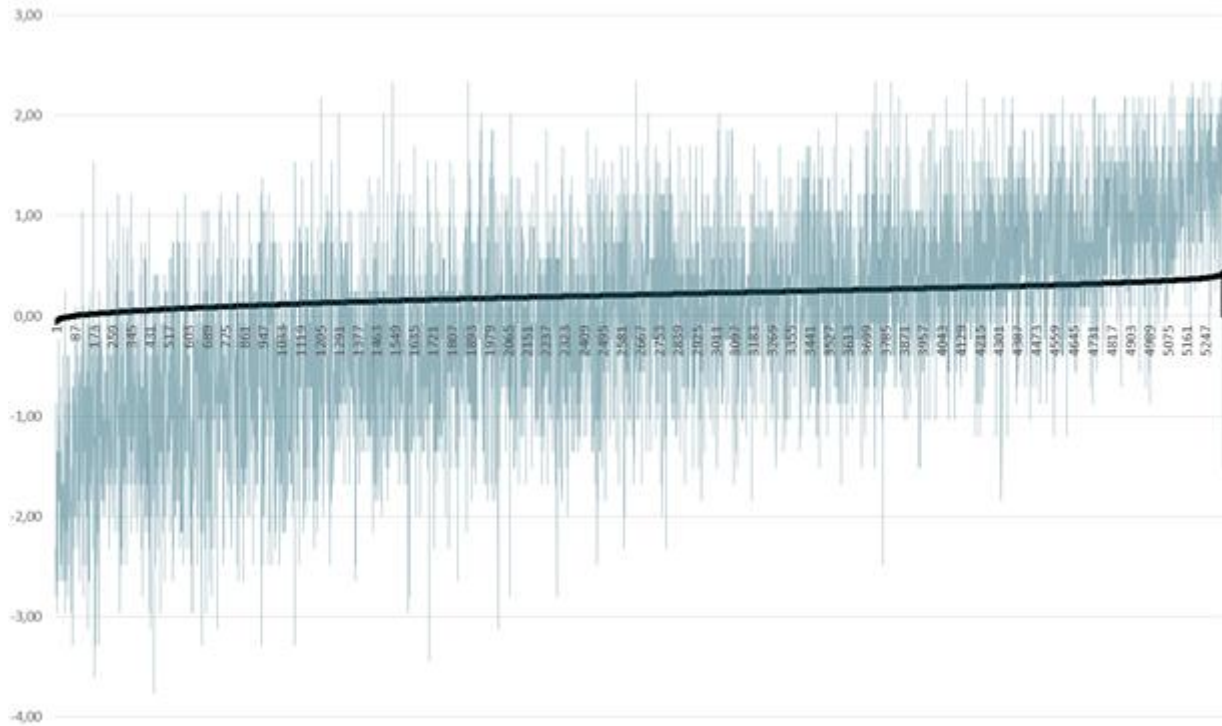
Figure 7 shows the opposite effect of the personality trait conscientiousness:

Figure 7. Individual respondents' co-product matrices sorted by their similarity with the sample correlation matrix (thick, straight line), plotted with their z-scores on Conscientiousness (thin jagged line). (N=5,331)

Dataset 4 is therefore particularly valuable in showing that people need not respond in ways that are semantically determined. Compliance with what is semantically expected seems to depend on a number of conditions, including attitude strength and personality traits.

Dataset number 1 originally contained a shortened version of the NEO that was not included in the previously published Arnulf et al.'s study (2014). Including information about personality in the present analysis allows us to check how demographics and personality may

affect semantic compliance in the MLQ data, which is usually where semantics show the strongest influence. We regressed the similarity of the co-products and distance matrices with the sample correlations and semantics on demographics, NEO scores, MLQ scale values and finally the respondents' management level. The results can be seen in Table 5. It appears that demographics are not important, but personality and satisfaction with own manager play a role for all matrices.

Table 5
Dataset 1, individual respondents' matrices explained by demographics, personality and leadership score

	Co-products				Distances			
	similarity with sample matrix		similarity with semantics		similarity with sample matrix		similarity with semantics	
Variables	Standard. beta	Sig.	Standard. beta	Sig.	Standard. beta	Sig.	Standard. beta	Sig.
(Constant)	-1.17	.24	-.80	.42	-1.25	.21	-1.44	.15
Birth year	1.13	.26	.85	.40	1.24	.22	1.42	.16
Sex	-3.37	.00	-2.80	.01	-1.14	.25	-1.11	.27
Education	.89	.37	.82	.41	-.27	.79	-.82	.41
Tenure	2.66	.01	1.89	.06	.02	.98	.23	.82
Neuroticism	.96	.34	.47	.64	-2.66	.01	-2.62	.01
Extraversion	-2.74	.01	-2.49	.01	.05	.96	.03	.98
Openness	.00	1.00	-.57	.57	2.17	.03	1.66	.10
Agreeableness	-1.80	.07	-1.35	.18	-2.03	.04	-1.66	.10
Conscientiousness	-.88	.38	-1.18	.24	2.41	.02	2.01	.04
Transformational leadership	45.56	.00	31.70	.00	14.72	.00	1.84	.00
Transactional leadership	-9.66	.00	1.25	.21	-.39	.69	3.82	.00
Laissez-faire	-35.53	.00	-39.48	.00	-13.75	.00	-15.12	.00
Management level	.53	.60	-2.03	.04	2.27	.02	.71	.48
Model adj. R^2	.91		.88		.58		.54	

It can thus be demonstrated that the co-product matrices capture diverse sources of variance influencing individuals' responses, such as their attitude strength, personality dispositions and management level. All of these variables seem to influence the attitude strength of individuals, and also their coherence with expected semantic properties of the scale.

But how much of this individual-level information is carried over into the sample characteristics? To explore this, we averaged the item co-products and item distances across respondents for each pair of items in all datasets. This level of analysis corresponds to the second step in formulas for correlations or co-variances, where sums of products are divided by sample size. Table 6 displays the relationships between the sample correlation matrix, the semantic matrix and the averaged co-products and distance matrices for all four samples. For all samples, the average distances matrix is more closely related to both the sample correlation matrix *and* the semantic matrix. Interestingly, in the first three datasets, the sample correlation matrix is also more strongly correlated with the semantic values than with any data generated by respondents (co-products or distances).

Table 6

Correlations between sample correlation matrix, semantic matrix and average co-product- and distances matrices

	DATASET 1 MLQ ONLY			DATASET 2 MLQ MOTIVES OUTCOMES		
	Sample correlation matrix	Semantic (MI) matrix	Average co-products matrix	Sample correlation matrix	Semantic (MI) matrix	Average co-products matrix
Semantic (MI) matrix	.81**			.70**		
Average co-products matrix	.83**	.71**		.40**	.50**	
Average distances matrix	.97**	.80**	.82**	.59**	.56**	.39**
	DATASET 3 DIV LEADERSHIP OUTCOMES			DATASET 4 NEO-FFI		
	Sample correlation matrix	Semantic (MI) matrix	Average co-products matrix	Sample correlation matrix	Semantic (MI) matrix	Average co-products matrix
Semantic (MI) matrix	.70**			.03		
Average co-products matrix	.40**	.50**		.39**	-.02	
Average distances matrix	.59**	.56**	.39**	.67**	.05*	.68**

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

In the table below, the sections are displaying the following analyses: Tables 7 and 8 explore the degree to which the co-products (Table 7) or distances (Table 8) are explained by semantics. As argued initially, for datasets 1 to 3, the item distance matrices are much closer to the semantic values than the co-products.

Tables 7 and 8

Average item distances and co-products predicted by semantics in linear regression

Table 7: Average co-products predicted by sematics in linear regression:

	DATASET 1			DATASET 2			DATASET 3			DATASET 4		
	Standard.		Adj.	Standard.		Adj.	Standard.		Adj.	Standard.		Adj.
	beta	Sig.	R2	beta	Sig.	R2	beta	Sig.	R2	beta	Sig.	R2
(Constant)		.000			.000			.000			.000	
MI values	.42	.000		.67	.000		.66	.000		-.06	.028	
NEWS cosines	.06	.428		-.45	.000		.19	.000		.15	.000	
BIZ cosines	-.08	.479		-.08	.373		-.09	.077		-.10	.000	
PR Cosines	.42	.000	.58	.32	.000	.26	-.07	.297	.47	-.06	.139	.02

Table 8: Average distances predicted by semantics in linear regression:

	DATASET 1			DATASET 2			DATASET 3			DATASET 4		
	Standard.		Adj.	Standard.		Adj.	Standard.		Adj.	Standard.		Adj.
	beta	Sig.	R2	beta	Sig.	R2	beta	Sig.	R2	beta	Sig.	R2
(Constant)		.000			.000			.000			.000	
MI values	-.44	.000		-.31	.000		-.84	.000		-.09	.001	
NEWS cosines	-.12	.042		-.64	.000		.17	.000		-.05	.189	
BIZ cosines	-.11	.191		-.16	.042		.12	.008		.02	.312	
PR Cosines	-.26	.001	.75	.52	.000	.35	-.25	.000	.65	.15	.000	.01

In Table 9, we use hierarchical regression to explore the degree to which the observed sample correlation matrix is differentially explained by semantics, co-products or distances. We entered the semantic values in step 1. It appears that the co-products have a minute impact on the observed sample correlations, as only the distances matrix scores improve the models in step 2. This also holds for dataset 4, the NEO scores.

Table 9: Semantics, avg. co-products and distances in hierarchical regression with sample correlation matrix as dependent variable:

Model		DATASET 1 Standard. beta	Sig.	Adj. R2	DATASET 2 Standard. beta	Sig.	Adj. R2	DATASET 3 Standard. beta	Sig.	Adj. R2	DATASET 4 Standard. beta	Sig.	Adj. R2
1	*(Constant)*		*.000*			*.000*			*.000*			*.174*	
	MI values	.43	.000		.64	.000		.71	.000		.03	.207	
	NEWS cosines	.17	.001		.70	.000		-.15	.001		-.08	.046	
	BIZ cosines	.06	.456		.21	.002		.00	.971		-.01	.791	
	PR Cosines	.30	.000	.79	-.83	.000	.53	.12	.057	.47	.06	.140	.00
2	*(Constant)*		.000			.000			.000			.000	
	MI values	.07	.000		.52	.000		.04	.036		-.04	.034	
	NEWS cosines	.08	.002		.58	.000		.04	.193		-.11	.001	
	BIZ cosines	-.02	.669		.18	.006		.09	.017		.00	.954	
	PR Cosines	.07	.047		-.73	.000		-.11	.013		.17	.000	
	AvgCoProds	**.09**	**.000**		**.07**	**.000**		**-.16**	**.000**		**-.13**	**.000**	
	AvgDist	*-.73*	*.000*	*.95*	*.23*	*.000*	*.57*	*-.91*	*.000*	*.75*	*-.76*	*.000*	*.46*

Finally, as displayed in Table 10 we regressed the average co-product and distance matrices on the semantic values, keeping the residuals, using these as estimates of scores after controlling for semantics. We then used hierarchical regression to predict the sample correlation matrices. Entering the residuals of the co-products in step 1, these still had predictive power until the residuals of the distance scores were entered in step 2. Here again, the co-products lost most of their predictive power, becoming insignificant in one case and of minute practical importance in the other cases.

Table 10: Co-products and distances controlled for semantics in stepwise regression with sample correlation matrix as dependent variable:

Model		DATASET 1 Standard. beta	Sig.	Adj. R2	DATASET 2 Standard. beta	Sig.	Adj. R2	DATASET 3 Standard. beta	Sig.	Adj. R2	DATASET 4 Standard. beta	Sig.	Adj. R2
1	(Constant)		.000			.000			.000			.000	
	Co-products controlled for semantics	.24	.000	.06	.09	.000	.01	.01	.635	.00	.39	.000	.15
2	(Constant)		.000			.000			.000			.000	
	Co-products controlled for semantics	.06	.080		.06	.002		.11	.000		.13	.000	
	Distances controlled for semantics	.37	.000	.16	.19	.000	.04	.54	.000	.27	.76	.000	.46

In this way, it appears that while the co-products carry substantial amounts of information about the individuals' response patterns (and their sources of variance), this information is largely lost in the average sample matrices. What appears instead is a function of the item distances, which are mainly informed by semantics.

Discussion

This study served two purposes: Primarily, we sought to develop a method for assessing how cognitive processes in the form of semantic structures influence individual survey respondents. Secondly, we wanted to show how the commonly applied statistical techniques departing from correlation or covariance matrices run a real risk of filtering out attitude strength, its own subject matter, to instead reflect the cognitive language parsing mechanisms of the respondents.

As there exists no previously validated method to determine this, we have argued that individual-level analyses can be performed by calculating two different types of individual-level

response matrices: the co-product and item distance matrices. Departing from the formula for product-moment correlation coefficients, we argue that by multiplying the paired responses for all individuals, the co-products should capture the response strength. Conversely, individual matrices computed as absolute distances (subtracting response X from response Y) should be relatively distinct from attitude strength and closer to the expected semantic values.

Our analysis seemed to support these assumptions. While the two kinds of matrices contain much of the same information and display similar distributions, the individual co-product matrices are most sensitive to the respondents' scale score levels, indicating their attitude strength. The distance matrices seem less sensitive to score level (attitude strength) and are more strongly related to the semantic properties of the scales.

The exploration of the individual co-product matrices show that these contain rich information about a number of respondent characteristics. Among the sources of variation found were personality, satisfaction with own manager and personal management level. In the case of the NEO-FFI, the personality trait levels exerted much stronger influence on the individual response patterns than the semantic properties of the items, in line with previous findings (Arnulf et al., 2014). However, when aggregated on sample levels, the explanatory power of the distance matrices dominated completely and there was almost nothing left of the information from the attitude strength. This was even true in the case of NEO-FFI scores, but in this case, the distances were still only weakly informed by the semantic properties of the items.

To understand how and why this happens, the NEO matrices were informative. The aggregated correlation matrix for the whole sample displayed a structure that most individuals do not display themselves. Even the mean co-product matrix for all respondents does not reach a

higher similarity than a correlation of .21 with the matrix for the whole sample, and no individual had a correlation higher than .45 with the sample structure. This means that for each individual, there is considerable variation in the way response patterns emerge and no one displays an entirely "normal" response structure in the statistical sense. This can be assumed to contribute to the predictive validity of the Big Five in many areas (van der Linden, te Nijenhuis, & Bakker, 2010) because individual response patterns are allowed to be widely different from the norm, hence capturing more idiosyncratic patterns characterizing the individuals. Responses to the items in NEO-FFI may reflect stable behavioral dispositions that are pervasive in the respondents' autobiographic memory. Hypothetically, personality may be a part of how the respondents apply their cognitive semantic processes, hence reflecting more "fundamental" psychological processes than those normally applied when answering surveys.

This is not so for the OB scales in the three other datasets, where the similarity is much higher. The mean correlation is around .45 and the maximum values as high as .90. This distribution may even have been a goal for the scholars developing the scales, as this will greatly enhance the alpha reliability, factor structure and model fit indices of the scales which are taken by many journals as quality criteria (Abdi, 2003; Arnulf et al., 2014; Fabrigar, Wegener, MacCallum, & Strahan, 1999; Nunnally & Bernstein, 2010). Failing to comply with such criteria has even been raised as an objection against the five-factor model (McKenzie, 1998).

When item semantics become the main common factor shaping the responses to the survey, it implies either that the respondents have little else in common, or that the survey elicits few other types of responses. In the NEO sample, Neuroticism seemed to make people answer in dissimilar ways that left little other than semantics in common. It also appears as if Conscientiousness increases the effort to comply with the requirements of the survey, and low

values of openness to experience seemed to create a propensity to answer the survey items in very literal ways, akin to the stale patterns detected by the computerized semantic algorithms.

The key to understanding the influence of semantics in surveys may reside in the well-known tenet of classical psychometrics, that all sources of variance in responses will be filtered out as noise if they are randomly distributed (Nunnally & Bernstein, 2010). In our study, this seems to happen with many sources of variation that determine attitude strength in surveys on OB. Even if attitude strength appears to be an important determinant of individual co-product matrices, reflecting meaningful differences among the respondents, the effects of attitude strength seem to disappear in the averaged sample matrices. What is left is the common denominator of most respondents – their compliance with the semantic criteria in the items.

It is compelling that the observed sample matrices, themselves computed by adding the co-products of items, seem to lose their roots in these values. The co-products should have emerged as important if only due to common method bias since these were the input values for the correlation matrix (P.M. Podsakoff et al., 2012b). When the matrix of average distances emerges as the most powerful predictor of the observed correlation matrix, it can only be explained by reducing the role of attitude strength to a mere carrier of semantic information.

These findings are troublesome for the interpretation of factor analyses and latent factor modelling on a theoretical level. Correlation as a method has always been known as being vulnerable to the problem of the "third variable" in whatever form (Antonakis, Bendahan, Jacquart, & Lalive, 2010). The present study suggests that the relationships picked up in correlation matrices may not even be due to extraneous factors influencing relationships. As recently shown by Maul (2017), data structures with good psychometric properties can be

obtained not only with nonsensical, but even with empty items that cannot possibly reflect any attitude strength whatsoever. From the semantic perspective in our case, it appears that the method of correlating Likert-scale data may easily become detached from its own subject-matter – the attitude strength – and instead pick up patterns created by the cognitive process of linguistic parsing when reading the items. Such a situation is clearly in conflict with Likert's original viewpoints on the value of his method (Likert, 1932), and shows how Thurstone was right in his doubts about the numbers collected by Likert's scales (Andrich, 1996).

When this happens, the method has lost its purported object and the findings are no longer concerned with the attitudes towards the contents of the survey. When factor structures and other model characteristics of survey data are based on semantics, they are no longer empirical, but semantic, similar to what Russell and Wittgenstein thought of as "logical" facts (Russell, 1922).

We believe that the reason for these important differences between the OB scales and the NEO personality items may be found in a subtle difference in the construction of items and scales. Psychometric techniques originate from intelligence testing (Likert, 1932 p 50) where the focus of interest was actual behaviors (e.g. response time or error rate) and not the endorsement of propositions. The items in personality tests such as the NEO sample observations of earlier behaviors, for example work habits, episodes of self-control, and aesthetic experiences. While the personality test items may superficially look similar to the OB counterparts, they are actually sampling a wide range of behavior frequencies. To the extent that respondents report overt and quantifiable behaviors, the information is non-semantic. Some authors argue that test items should include broad samples of behavior instead of seeking high alpha reliabilities through similarly worded items (Kline, 2000). This approach probably creates messier factor structures

and alphas, but it does seem to cluster non-semantic relationships between behaviors in peoples' lives.

The OB items are much closer to Likert's original view of his instruments as "dispositions towards overt action" or an "indirect method" (Likert, 1932 p. 9). Following Likert's work, measures of attitudes have been asking the respondents to rate opinions on anything from US-Japanese relations (Likert, 1932) to ease of use in computer software (Davis, 1989). Such questions rarely tap information about observed behaviors, but are much more likely to address purely cognitive constructions (Gefen & Larsen, 2017). As shown by Gollwitzer & Sheeran (2006), attitudinal measures can only modestly predict behavior unless the subject has previous observations of own behavior. Researchers may infuse instruments with semantic relationships that will raise alpha reliabilities, but this will not necessarily improve the predictive validity of the measures.

The application of statistical methods to the purported measurements of latent, but verbally defined constructs is mathematically complex and fraught with unsolved questions about quantification (Slaney, 2017). To the degree that items tap information about more or less overt behaviors, their quantification may be more straightforward than measurements of attitudes. We are currently working on designs to explore the relationships between semantics and behavioral assessments in more detail. On the other hand, we think this opens up a fascinating new perspective on Likert scales as a psychological method. If the semantic influence on survey responses has been both pervasive and neglected for decades, it means that the language parsing mechanisms of the brain have been taken almost for granted. The capability of a fluent reader to parse and understand language is usually quite effortless (Kahneman, 2011), and at the same time remarkably precise in its ability to recognize minute differences between short items in surveys

(Michell, 1994). An improved theory of cognitive processes involved in survey responses –

STSR – offers at least two interesting further paths.

First, the cognitive structure needs to be consistent across subjects, because this is a

prerequisite to hold different opinions about the same subject matter (Coombs & Kao, 1960;

Russell, 1922). It would be interesting to establish a method to describe the common cognitive

model that respondents need to share in order to respond in reliably different ways. The common

ways of using Likert-scale data until now has largely been to find statistically significant patterns

using null hypothesis significance testing (Krueger, 2001). This may be one of the reasons why

models based on survey data suffer from common method variance and poor predictive validity

(Bagozzi, 2011; McClelland, Koestner, & Weinberger, 1989; P.M. Podsakoff et al., 2012b). If we

had a way of establishing an expected semantic pattern of responses, it may serve as a better

platform for testing non-obvious behavioral relationships than simply rejecting a null hypothesis

(Arnulf, Larsen, Martinsen, & Bong, 2015a; Rodgers, 2010; Smedslund, 1994).

Secondly, and conjunctive to this, it may be much more interesting and informative to

explore how survey data deviate from the semantically expected. Our capacity for abstract

thinking seems to originate from the brain's spatial capabilities, and these spatial representations

are always distorted (Tversky, 1993). It is conceivable that individual ideosyncracies may distort

the semantic spaces of individuals also, and that these differences are valuable psychological

information. The findings of the present study show that respondents leave multiple types of

information in their response patterns. If this information is reduced to the semantic

commonalities of the participants, we may miss the true value of the collected information. A

better understanding of the semantic structures and how they are bent by respondents may offer

new insights with better predictive validity because they are closer to the characteristics of the respondents.

Two examples may be useful here: The semantic algorithms predict that transformational leadership will be linked to intrinsic motivation and work outcomes simply because the items have overlapping meanings. A person who endorses his boss (transformational leadership) and likes his tasks (intrinsically motivated) will also not look for a job (turnover intention) because this is simply what these sentences imply. These connections will be discernible in a large sample of people simply because this is how they read the items, on average. It might be more interesting to know if or how people with strong positive or negative attitudes towards their superiors or jobs also bend their cognitions to fit their world views. So far, we have had no agreements on methods to explore such relationships.

Limitations and Implications for Further Research

This is a type of analysis that has never been undertaken before. Thus, we have few or no criteria for assessing the quality of the displayed statistics. Also, the four samples fall into two very distinct groups – while the OB scales seem strongly determined by semantics, this does not seem to be the case for the personality scales. This opens up a possibility that different types of Likert scales may display very different properties with semantics. This selection of data samples may have influenced or biased our findings, but we believe that the present study has opened a path to more direct investigations of this topic.

An important set of limitations come from the semantic algorithms. As anyone who has tried automatic translation tools will know, text algorithms are not perfect. The MI algorithm has no context-specific knowledge, but is more suitable to short texts like survey items. The LSA

cosines are conversely dependent on the semantic spaces used and the number of dimensions applied (usually denoted k). We simply do not possess knowledge about how variations in these characteristics shape the statistics we report here.

Further, we have chosen to only analyze data from the most commonly used types of responses scales where respondents are asked to indicate their attitudes in ways that are later quantified along a dimension. There are many more types of scales (e.g., Brennan, 2010; Michell, 2013; van Schuur, 2003) that are likely to display other relationships with semantics.

Also, none of our analyses considered how responses were nested within surveys, departments or organizations. We are currently preparing multi-source-multi-method studies of new datasets to explore the effect of more complex analyses. We therefore call for further research along these lines. It would be interesting to have more incisive studies of the statistics we introduce here. Further exploration of the statistics of individual response patterns and the influence of semantics is necessary to establish a more solid foundation for further empirical analyses.

With the rapidly increasing use of big data and other analytical techniques, the role of surveys is changing (Kern et al., 2014; M. Kosinski, Bachrach, Kasneci, Van-Gael, & Graepel, 2012; Michal Kosinski, Stillwell, & Graepel, 2013). The study of how computer algorithms can be applied to psychological materials is a promising field for psychological methodology, already applied in practice by a series of technological companies (Markowetz, Blaszkiewicz, Montag, Switala, & Schlaepfer, 2014). Another line of future research would be to explore how these findings generalize to other domains of Likert-scale data. A more nuanced series of findings could help establish a more solid theoretical foundation for the STSR.

Our final goal would be the establishment of a general model for the semantic processes involved in survey responses. Attempts at this or calls for such endeavors have been published intermittently over the years (Feldman & Lynch, 1988; Michell, 1994, 2013; Schwarz, 1999). With the increasing capability of digital tools (Landauer, 2007; Larsen & Bong, 2016) and neurobiological methods in mapping semantic understanding (Honey et al., 2012), the time seems ripe for a better understanding of the actual psychological processes that end up in the form of scores on Likert scales.

References

Abdi, H. (2003). Factor rotations in factor analysis. In M. Lewis-Beck, A. Bryman, & T. Futing (Eds.), *Encyclopedia of social sciences research methods* (pp. 792-795). Thousand Oaks, CA: Sage.

Aiken, L. S., West, S. G., & Reno, R. R. (1991). *Multiple regression: Testing and interpreting interactions*: Thousand Oaks, CA: Sage.

Andrich, D. (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. *British Journal of Mathematical & Statistical Psychology, 49*, 347-365.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly, 21*(6), 1086-1120.

APA (American Psychological Association), (2009). History of the standards. *Standards for educational and psychological testing.*

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, 2*(e3), 1-28. doi:http://dx.doi.org/10.15714/scandpsychol.2.e3

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., & Bong, C. H. (2014). Predicting survey responses: How and why semantics shape survey statistics in organizational behavior. *PLoS ONE, 9*(9), 1-13. doi:10.1371/journal.pone.0106361

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., & Bong, C. H. (2015b). A new approach to psychological measures in leadership research. *Scandinavian Psychologist, 2*, e4. doi:http://dx.doi.org/10.15714/scandpsychol.2.e4

Avolio, B. J., Bass, B. M., & Jung, D. I. (1995). *Multifactor Leadership Questionnaire technical report.* Redwood City, CA.: Mind Garden.

Bagozzi, R. P. (2011). Measurement and meaning in information systems and organizational research: Methodological and philosophical foundations. *MIS Quarterly, 35*(2), 261-292.

Bernerth, J. B., Armenakis, A. A., Feild, H. S., Giles, W. F., & Walker, H. J. (2007). Leader–member social exchange (LMSX): Development and validation of a scale. *Journal of Organizational Behavior, 28*(8), 979-1003. doi:10.1002/job.443

Borsboom, D. (2008). Latent variable theory. *Measurement, 6*, 25-53.

Borsboom, D. (2009). Educational measurement: Book review. *Structural Equation Modeling, 16*, 702-711. doi:10.1080/10705510903206097

Brennan, R. L. (2010). Generalizability theory. In P. P. B. McGaw (Ed.), *International encyclopedia of education (Third Edition)* (pp. 61-68). Oxford: Elsevier.

Cammann, C., Fichman, M., Jenkins, D., & Klesh, J. (1979). *The Michigan organizational assessment questionnaire.* University of Michigan, Ann Arbor: University of Michigan.

Cascio, W. F. (2012). Methodological issues in international HR management research. *International Journal of Human Resource Management, 23*(12), 2532-2545. doi:10.1080/09585192.2011.561242

Conger, J. A., & Kanungo, R. N. (1998). *Charismatic leadership in organizations*. Thousand Oaks, CA.: Sage.

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

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO personality Inventory: Professional manual*. Odessa, Fl: Psychological Assessment Resources.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

de Vries, J., Byrne, M., & Kehoe, E. (2015). Cognitive dissonance induction in everyday life: An fMRI study. *Social Neuroscience, 10*(3), 268-281. doi:10.1080/17470919.2014.990990

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, 13(3), 319-340. doi:10.2307/249008

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science, 41*(391-407).

Dennis, S., Landauer, T., Kintsch, W., & Quesada, J. (2003). *Introduction to latent semantic analysis*. Paper presented at the 25th Annual Meeting of the Cognitive Science Society, July 31st – Aug. 2nd. Boston, Mass.

Dennis, S., Landauer, T., Kintsch, W., & Quesada, J. (2013). *Introduction to latent semantic analysis*. University of Colorado: Boulder, Co.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272-299.

Feldman, J. M., & Lynch, J. G. J. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of Applied Psychology, 73*(3), 421-435.

Felfe, J., & Schyns, B. (2014). Romance of leadership and motivation to lead. *Journal of Managerial Psychology, 29*(7), 850-865. doi:10.1108/jmp-03-2012-0076

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research*. Reading, Mass.: Addison-Wesley.

Fisher, H. E., Island, H. D., Rich, J., Marchalik, D., & Brown, L. L. (2015). Four broad temperament dimensions: Description, convergent validation correlations, and comparison with the Big Five. *Frontiers in Psychology, 6*, 18. doi:10.3389/fpsyg.2015.01098

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*, 39-50.

Gefen, D., & Larsen, K. R. (2017). Controlling for lexical closeness in survey research: A demonstration on the Technology Acceptance Model. *Journal of the Association for Information Systems* (forthcoming).

Gollwitzer, P., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in Experimental Social Psychology, 38*, 69-118.

Honey, C. J., Thomson, C. R., Lerner, Y., & Hasson, U. (2012). Not lost in translation: Neural responses shared across languages. *Journal of Neuroscience, 32*(44), 15277-15283.

Ilies, R., Nahrgang, J. D., & Morgeson, F. P. (2007). Leader-member exchange and citizenship behaviors: A meta-analysis. *Journal of Applied Psychology, 92*(1), 269-277.

Jackson, R. L., Hoffman, P., Pobric, G., & Ralph, M. A. L. (2016). The semantic network at work and rest: Differential connectivity of anterior temporal lobe subregions. *Journal of Neuroscience, 36*(5), 1490-1501. doi:10.1523/jneurosci.2999-15.2016

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294-316). Newbury Park: Sage.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Chicago, IL: SSI Scientific Software.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Park, G., Ungar, L. H., Stillwell, D. J., . . . Seligman, M. E. P. (2014). From "sooo excited!!!" to "so proud": Using language to study development. *Developmental Psychology, 50*(1), 178-188. doi:10.1037/a0035048

Kim, S., Hahn, H. J., & Lee, J. (2015). Organizational attitudes as precursors to training performance. *Human Resource Development Quarterly, 26*(4), 409-429. doi:10.1002/hrdq.21218

Kline, P. (2000). *Handbook of psychological testing* (2nd ed.). London: Routledge.

Kosinski, M., Bachrach, Y., Kasneci, G., Van-Gael, J., & Graepel, T. (2012). Crowd IQ: measuring the intelligence of crowdsourcing platforms. *Proceedings of the 4th Annual ACM Web Science Conference*, *Evanston, Illinois, June 22 - 24, 151-160*. doi: 10.1145/2380718.2380739

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences, 110*(15), 5802-5805. doi:10.1073/pnas.1218772110

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist, 56*, 16-26.

Kuvaas, B. (2006a). Performance appraisal satisfaction and employee outcomes: Mediating and moderating roles of motivation. *The International Journal of Human Resource Management, 17*(3), 504-522.

Kuvaas, B. (2006b). Work performance, affective commitment, and work motivation: The roles of pay administration and pay level. *Journal of Organizational Behavior, 27*(3), 365-385.

Kuvaas, B., Arnulf, J. K., & Dysvik, A. (2008). *Social and economic exchange perceptions and intrinsic motivation among knowledge workers*. Paper presented at the Academy of Management Annual Meeting, Anaheim, California.

Lamiell, J. T. (2013). Statisticism in personality psychologists' use of trait constructs: What is it? How was it contracted? Is there a cure? *New Ideas in Psychology, 31*(1), 65-71. doi:http://dx.doi.org/10.1016/j.newideapsych.2011.02.009

Landauer, T. K. (2007). LSA as a theory of meaning. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsh (Eds.), *Handbook of Latent Semantic Analysis* (pp. 3-34). Mahwah, New Jersey: Larence Erlbaum Associates, Publishers.

LaPiere, R. T. (1934). Attitudes vs. actions. *Social Forces, 13*, 230-237.

LaPointe, L. L. (2013). *Paul Broca and the origins of language in the brain*. San Diego, CA US: Plural Publishing.

Larsen, K. R., & Bong, C. H. (2013) Construct identity detection. *Leeds School Working Paper Series* (pp. 1-70). Boulder, CO: University of Colorado, Boulder.

Larsen, K. R., & Bong, C. H. (2016). A tool for addressing construct identity in literature reviews and meta-analyses. *MIS Quarterly, 40*(3), 529-551.

Larsen, K. R., & Monarchi, D. E. (2004). A mathematical approach to categorization and labeling of qualitative data: The latent categorization method. *Sociological Methodology, 34*(1), 349-392.

Larsen, K. R., Nevo, D., & Rich, E. (2008). *Exploring the semantic validity of questionnaire scales*. Paper presented at the Proceedings of the 41st Hawaii International Conference on System Sciences, Waikoloa, Hawaii.

Li, Y., McLean, D., Bandar, Z. A., O'Shea, J. D., & Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering, 18*(8), 1138-1150.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 1-55.

Lynam, D. R., Hoyle, R. H., & Newman, J. P. (2006). The perils of partialling. *Assessment, 13*(3), 328-341. doi:doi:10.1177/1073191106290562

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly, 35*(2), 293-334.

Markowetz, A., Blaszkiewicz, K., Montag, C., Switala, C., & Schlaepfer, T. E. (2014). Psycho-Informatics: Big data shaping modern psychometrics. *Medical Hypotheses, 82*(4), 405-411. doi:10.1016/j.mehy.2013.11.030

Marsh, H. W., Ludtke, O., Muthen, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment, 22*(3), 471-491. doi:10.1037/a0019227

Martinsen, Ø. L., Nordvik, H., & Østbø, L. (2003). *Norske versjoner av NEO PI - R og NEO FFI (Norwegian versions of the NEO PI-R and NEO FFI)*. Oslo: Gyldendal Akademisk forlag.

Maul, A. (2017). Rethinking Traditional Methods of Survey Validation. *Measurement: Interdisciplinary research and perspectives*, 1-19. doi:10.1080/15366367.2017.1348108

McClelland, D. C., Koestner, R., & Weinberger, J. (1989). How do self-attributed and implicit motives differ. *Psychological Review, 96*(4), 690-702.

McKenzie, J. (1998). Fundamental flaws in the five factor model: A re-analysis of the seminal correlation matrix from which the "openness-to-experience" factor was extracted. *Personality and Individual Differences, 24*(4), 475-480. doi:10.1016/s0191-8869(97)00223-7

Meyer, J. P., Allen, N. J., & Smith, C. A. (1993). Commitment to organizations and occupations: Extension and test of a three-component conceptualization. *Journal of Applied Psychology, 78*(4), 538-551.

Michell, J. (1994). Measuring dimensions of belief by unidimensional unfolding. *Journal of Mathematical Psychology, 38*(2), 244-273.

Michell, J. (2013). Constructs, inferences, and mental measurement. *New Ideas in Psychology, 31*(1), 13-21. doi:http://dx.doi.org/10.1016/j.newideapsych.2011.02.004

Mihalcea, R., Corley, C., & Strapparava, C. (2006a). *Corpus-based and knowledge-based measures of text semantic similarity.* Paper presented at the 21st national conference on Artificial intelligence in Boston, Mass.

Mihalcea, R., Corley, C., & Strapparava, C. (2006b). Corpus-based and knowledge-based measures of text semantic similarity. *AAAI, 6*, 775-780.

Miller, G. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*(11), 39-41. doi:http://dx.doi.org/10.1145/219717.219748

Mohler, M., & Mihalcea, R. (2009). *Text-to-text semantic similarity for automatic short answer grading.* Paper presented at the 12th Conference European Chapter of the Association for Computational Linguistics (EACL 2009), Athens, Greece.

Nakagawa, J., Takahashi, M., Okada, R., Matsushima, E., & Matsuda, T. (2015). Women's preference for a male acquaintance enhances social reward processing of material goods in the anterior cingulate cortex. *PLoS ONE, 10*(8), 14. doi:10.1371/journal.pone.0136168

Nimon, K., Shuck, B., & Zigarmi, D. (2015). Construct overlap between employee engagement and job satisfaction: A function of semantic equivalence? *Journal of Happiness Studies*, *17*(3), 1149-1171. doi:10.1007/s10902-015-9636-6

Nunnally, J. C., & Bernstein, I. H. (2010). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012a). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 65*, 539-569.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012b). Sources of method bias in social science research and recommendations on how to control it. In S. T. Fiske, D. L. Schacter, & S. E. Taylor (Eds.), *Annual Review of Psychology* (Vol. 63, pp. 539-569). Palo Alto: Annual Reviews.

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

Poli, R., Healy, M., & Kameas, A. (2010). WordNet. In C. Fellbaum (Ed.), *Theory and applications of ontology: Computer applications* (pp. 231-243). New York: Springer.

Prinz, W., Aschersleben, G., & Koch, I. (2009). Cognition and action. In E. Morsella, J. A. Bargh, & P. M. Gollwitzer (Eds.), *Oxford handbook of human action*. Boston: Oxford University Press, Inc.

Regev, M., Honey, U., & Hasson, U. (2013). Modality-selective and modality-invariant neural responses to spoken and written narratives. *Journal of Neuroscience, 33*(40), 15978-15988.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist, 65*(1), 1-12.

Russell, B. (1922). An introduction to the tractatus logico-philosophicus. In L. Wittgenstein (Ed.), *Tractatus logico-philosophicus*. London: Kegan Paul.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93-105.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shore, L. M., Tetrick, L. E., Lynch, P., & Barksdale, K. (2006). Social and economic exchange: Construct development and validation. *Journal of Applied Social Psychology, 36*(4), 837-867.

Slaney, K. L., & Racine, T. P. (2013). What's in a name? Psychology's ever evasive construct. *New Ideas in Psychology, 31*(1), 4-12.

doi:http://dx.doi.org/10.1016/j.newideapsych.2011.02.003

Slaney, K. L. (2017). *Validating Psychological Constructs : Historical, Philosophical, and Practical Dimensions*. London, UK: Palgrave MacMillan.

Smedslund, J. (1994). Nonempirical and empirical components in the hypotheses of 5 social-psychological experiments. *Scandinavian Journal of Psychology, 35*(1), 1-15.

Stogdill, R. M. (1963). *Manual for the leader behavior description questionnaire, form XII*. Columbus, OH: Bureau of Business Research, Ohio State University.

Triandis, H. C. (1980). *Value, attitudes and interpersonal behavior*. In M. M. Page (Ed.), Nebraska Symposium on Motivation, Beliefs, Attitudes and Values, Lincoln, NE: University of Nebraska.

Tversky, B. (1993). Cognitive maps, cognitive collages, and spatial mental models. In A. U. Frank & I. Campari (Eds.), *Spatial information theory: A theoretical basis for GIS* (pp. 14-24). Berlin: Springer.

Van Berkum, J. J. A., Koornneef, A. W., Otten, M., & Neuwland, M. S. (2007). Establishing reference in language comprehension: An electrophysiological perspective. *Brain Research, 1146*, 158-171. doi:10.1016/j.brainres.2006.06.091

van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality, 44*, 315-327. doi:10.1016/j.jrp.2010.03.003.

Van Dyne, L., & LePine, J. A. (1998). Helping and voice extra-role behaviors: Evidence of construct and predictive validity. *Academy of Management Journal, 41*(1), 108-119.

Van Knippenberg, D., & Sitkin, S. B. (2013). A critical assessment of charismatic-transformational leadership research: Back to the drawing board? *The Academy of Management Annals, 7*(1), 1-60. doi:http://dx.doi.org/10.1080/19416520.2013.759433

van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis, 11*(2), 139-163. doi:10.1093/pan/mpg002

Wittgenstein, L. (1922). *Tractatus logico-philosophicus*. London: Kegan Paul.

Zimbardo, P. G., Ebbesen, E. B., & Maslach, C. (1977). *Influencing attitudes and changing behavior* (2nd ed.). Reading, MA: Addison Wesley.