Predicting Survey Responses: How and Why Semantics Shape Survey Statistics on Organizational Behaviour

Jan Ketil Arnulf[1]*, Kai Rune Larsen[2], Øyvind Lund Martinsen[1], Chih How Bong[3]

1 Department of Leadership and Organizational Behaviour, BI Norwegian Business School, Oslo, Norway, **2** Management and Entrepreneurship Division, Leeds School of Business, University of Colorado at Boulder, Boulder, Colorado, United States of America, **3** Faculty of Computer Science and Information Technology, University of Malaysia at Sarawak, Sarawak, Malaysia

Abstract

Some disciplines in the social sciences rely heavily on collecting survey responses to detect empirical relationships among variables. We explored whether these relationships were *a priori* predictable from the semantic properties of the survey items, using language processing algorithms which are now available as new research methods. Language processing algorithms were used to calculate the semantic similarity among all items in state-of-the-art surveys from Organisational Behaviour research. These surveys covered areas such as transformational leadership, work motivation and work outcomes. This information was used to explain and predict the response patterns from real subjects. Semantic algorithms explained 60–86% of the variance in the response patterns and allowed remarkably precise prediction of survey responses from humans, except in a personality test. Even the relationships between independent and their purported dependent variables were accurately predicted. This raises concern about the empirical nature of data collected through some surveys if results are already given *a priori* through the way subjects are being asked. Survey response patterns seem heavily determined by semantics. Language algorithms may suggest these prior to administering a survey. This study suggests that semantic algorithms are becoming new tools for the social sciences, opening perspectives on survey responses that prevalent psychometric theory cannot explain.

Citation: Arnulf JK, Larsen KR, Martinsen ØL, Bong CH (2014) Predicting Survey Responses: How and Why Semantics Shape Survey Statistics on Organizational Behaviour. PLoS ONE 9(9): e106361. doi:10.1371/journal.pone.0106361

Editor: Ramin Homayouni, University of Memphis, United States of America

Received March 26, 2014; **Accepted** August 6, 2014; **Published** September 3, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All data are unrestrictedly available on personal contact with the corresponding author: jan.k.arnulf@bi.no. Due to copyright issues affecting parts of the surveys used (the MLQ), we can only share the semantic values and the survey responses, but not the survey item wordings, which are the property of Mindgarden Inc. However, the items are purchasable from Mindgarden.com, and we used the form "MLQ 360 Form 5X Short". Any interested researchers should be able to use our data freely, but need to take proper responsibility in their use of the MLQ items. We will therefore avoid a publicly accessible repository, but promise to offer interested parties all information necessary to reproduce our findings or apply them to other materials. The prerequisite is a clarifying dialogue with the corresponding author.

Funding: The authors thank the United States National Science Foundation for research support under grant NSF 0965338 and the National Institutes of Health through Colorado Clinical and Translational Sciences Institute for research support under NIH/CTSI 5 UL1 RR025780. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: jan.k.arnulf@bi.no

Introduction

In this study, we explore how survey response patterns may be predicted using information available prior to conducting a survey. Such techniques have several interesting consequences for theory development and testing in the social sciences.

Many social science disciplines acquire data from surveys. The focus of interest is usually in how different variables relate to each other, allowing exploration of relationships such as those between leadership, motivation and work outcomes. To understand how these variables are related, researchers have hypothesised the existence of 'latent variables' – hidden sources of quantitative variation stemming from variables such as different types of leadership and motivation [1].

The Achilles heel of this research is the nature of variation in survey scores. The most common input to the computational tools is the inter-item correlation matrix, or the degree to which any two items in the survey tend to co-vary in a systematic way [2]. Commonly, the non-random patterns in survey responses are understood to reflect the systematic influence of some psychological or social variables on the respondents.

However, a fundamentally different explanation is possible. The main source of quantitative variation in the surveys may instead be the degree of semantic overlap among the items. We will attempt to show empirically how a *semantic theory of survey response* (STSR) allows an alternative interpretation of survey data from areas such as leadership, motivation and self-reported work outcomes, affecting views on theory formation, research methods and empirical data.

A Theory of Semantic Survey Response

Empirical, psychological, and semantic components of variance in survey data

The statistical treatment of survey data in the social sciences has developed as a discipline often referred to as 'psychometrics', originally developed from research on intelligence [3,4]. Intelligence tests consist of (often non-verbal) tasks to be solved, and responses are recorded fairly objectively as ratings of error frequency or response speed, and are therefore not susceptible to semantically determined responses. Later, Rensis Likert introduced a method familiar to most people today – having respondents rate a statement on a scale from "strongly approve" to "strongly disapprove" or similar [5]. Seemingly akin to intelligence tests, this is something altogether different and the origins and nature of the recorded variance are debatable [6–8].

We cannot know *a priori* how a respondent will rate a given item, e.g. "I like to work here". But once the respondent has chosen a value, the values for the next items may probably be given to some extent. To take an example: "Today is Monday". Someone rating this as "very true" is very likely to give the same rating to "Tomorrow is Tuesday". Most items are not as obviously linked. But someone affirming that "I like to work here" may with a similar probability endorse "I do not want to quit this job".

This semantic linkage of items is the core of what we believe to be a misunderstanding in survey-based research, demonstrable through semantic research. General psychometric theory asserts that some semantic overlap is necessary to create intra-scale consistency, usually measured by the formula called 'Cronbach's alpha' [1]. But the semantic overlap needs to stop there. If the semantic overlap continues across scales, it is regarded as a contamination of the data since one scale will automatically correlate with another. To prevent this, prevalent psychometric practices call for statistical procedures called exploratory and confirmatory factor analysis (CFA). By convention, the proper conduction of such analyses is taken as proof that relationships among variables are empirical and not self-evident [9].

As we will show empirically, this assumption does not hold. The semantic relationships hold across different scales despite their apparent separation by factor analysis. The resulting inter-item correlations can be explained by their semantic relationships. This is unfortunate because it undermines the value of factor analysis in establishing scale independency and also raises fundamental questions about the empirical object of such techniques.

Our concerns are not new in research on surveys and psychometric theory. More than five decades ago, Coombs and Kao [10] demonstrated that factor analysis in itself will always produce an extra factor that they called the "social utility function". This factor determines the data structure simply due to the meaning of the items, which all respondents would need to interpret in order to answer the survey. Coombs developed this function into a psychometric theory called "unidimensional unfolding". As Coombs predicted, this has been shown to influence factor analyses [11,12]. More importantly, experiments have shown that the quantitative properties of surveys are created by the semantic properties of items and their answering categories [8]. This may explain how independent research has shown respondents to provide responses where they in reality hold no opinion, or even to totally fictitious topics [6,7].

The need of the digital community to store, search, index and extract large amount of texts has stimulated the development of techniques that are sufficiently reliable and developed to take on survey research [13]. The task at hand is theoretically straightforward: If the overlap of meaning between any two survey items can be estimated quantitatively, the estimate can be used to explore the degree to which respondents are actually answering according to what is semantically expected.

We have chosen two types of text algorithms for this task. One is called *latent semantic analysis* (LSA), which has previously been shown to perform very similarly to human language learning using large chunks of text as its input. The second type of algorithm is corpus-based, which means that it uses a lexical database and knowledge about sentence syntax structure as input. The one we use here will be referred to as 'MI', a term used by its developers (MI is just a name for the algorithm) [14,15]. Both types of algorithm explore the semantic similarity of two different texts and return a measure expressing probable degree of semantic overlap. The team of authors has access to more advanced and efficacious techniques, but LSA and MI are used because they have been previously published, are well understood, allow easy replication, and remove uniqueness of algorithms as an explanation for our findings. We will refer to these two techniques together as *semantic analyses* and their numerical output as *semantic similarity indices*.

LSA functions by analysing texts to create a high-dimensional 'semantic space' in which all terms have specific locations, represented as vectors. LSA can then 'understand' new texts as combinations of term vectors in this space. LSA aggregates the word contexts in which a given word does/does not appear and provides a set of constraints that determines the similarity of meanings of words and sets of words. Thus, when two terms occur in contexts of similar meaning –even in cases where they never occur in the same passage –the reduced-dimension solution represents them as similar. Similarity is indicated by the cosines of the vectors in semantic space, taking on values between −1 and 1. Some practical examples: The two sentences "doctors operate on patients" and "physicians do surgery" have no words in common, but a commonly used LSA semantic space called TASA (Touchstone Applied Science Associates) estimates their overlap in meaning at .80. Furthermore, sentences with similar words do not necessarily appear as similar. For example, the LSA cosine for the two expressions "the radius of spheres" and "a circle's diameter" is .55, but the cosine for the sentence pair "the radius of spheres" and "the music of spheres" is only .01 [16].

LSA represents a sparse matrix of documents (columns) vs. terms-in-those-documents (rows). The matrix is generally set to downweigh common words. It is sometimes normalized before using an algorithm –singular value decomposition –similar to factor analysis. LSA then yields the aforementioned semantic space. This method now has well-documented text-recognition applications [17,18]. LSA works across languages. It is viable in both research and commercial contexts, and it performs almost as well as humans on complex knowledge-management and integration tests [19]. The usefulness of this technique has been documented in determining identities of a wide range of constructs in the Information Systems discipline [13,20].

Our approach was to let LSA detect accumulated knowledge and semantic relationships within texts relevant to respondents of organisational surveys. We defined relevant texts as articles from three different domains of media: Business-press texts, general newspaper texts, and PR-related texts.

The business-press texts were excerpts from *The Wall Street Journal*, *Business Week*, *Forbes* and *Fortune*. These excerpts covered a total of 84,836 texts from the years 1998–2007, covering a total of 45,816,686 words with 169,235 unique words.

The news excerpts were from *The New York Times*, *Los Angeles Times*, *Chicago Tribune*, *The Washington Post*, *The Boston Globe*, *USA Today*, *Houston Chronicle*, *San Francisco Chronicle* and *The Denver Post*. The years covered were again 1998–2007, including 162,929 texts covering 107,239,064 total words with 286,312 unique words.

The PR statements were taken from *PR Newswire*, covering the years 2003–2007. This sample included 212,484 texts with 151,450,055 total words and 423,001 unique words.

These materials allowed us to create three distinct 'semantic spaces', i.e. high-dimensional spaces in which all terms have a specific vector or location, allowing LSA to 'understand' the text of survey items. Every survey item in the study was projected into