Exploring the Semantic Validity of Questionnaire Scales

Kai R. Larsen
Leeds School of Business
University of Colorado, Boulder
kai.larsen@colorado.edu

Dorit Nevo
Schulich School of Business
York University
dnevo@schulich.yorku.ca

Eliot Rich
School of Business
University at Albany
e.rich@albany.edu

Abstract

Many behavioral researchers have been or are currently engaged in survey research, analyzing results using statistical methods. Respondents are often asked to fill out questionnaires leading to questionnaire fatigue and reluctance to conscientiously respond. Furthermore, in spite of the popularity of the approach, serious unanswered questions remain about what questionnaires actually measure. To answer these questions, this paper ventures into a new area of inquiry within survey research, providing a semantic analysis of questionnaires. In so doing we diverge from traditional survey validity measures, and offer a cutting edge approach to questionnaire validation with important contributions to future research[1].

1. Introduction

Many behavioral researchers have been or are currently engaged in survey research generally analyzing results using statistical methods. Because such approaches are ubiquitous, respondents often experience questionnaire fatigue leading to unusable data [1]. Furthermore, in spite of the popularity of the approach, serious unanswered questions remain about what questionnaires actually measure. Is it possible that when fatigue and unwillingness to conscientiously respond set in, questionnaires measure something else? If so, what do the questionnaires measure in such settings? To answer these questions, this paper ventures into a new area of inquiry within survey research, focusing on the semantic analysis of questionnaires. In so doing we diverge from traditional survey validity measures, and offer a cutting edge approach to questionnaire validation with important contributions to future research.

[1] This research was partially funded by a grant from the Social Sciences and Humanities Research Council of Canada (SSHRC)

Current practices of survey research engage a set of measures that enable researchers to assess the validity and reliability of questionnaires. Measurement scales created for surveys are assessed by field experts to determine their domain coverage (content validity) and comprehensibility (face validity). Once responses are received, various statistical methods are employed to ensure that items within each measurement scale are closely correlated (reliability) and that the scales measure what they purport to measure and are sufficiently differentiated from those measuring other constructs (construct validity). However, one important aspect of questionnaire validity – namely, language selection – is still underdeveloped. The specific words selected and combined into a questionnaire statement are of paramount importance, and as different words are stringed together, the meaning of individual words change to account for the new context provided by additional words. In fact, as respondents with similar experiences or attitudes respond to a questionnaire, high correlations between questionnaire statements emerge. Recently, linguists have found that our shared world view and accumulated knowledge may be captured in our language [2]. This presents a troubling conundrum for questionnaire researchers. How, then, can we distinguish between standard variance from shared language itself and variance from attitudes or beliefs? Current measures of face and content validity are highly subjective, and the statistical tests of reliability and construct validity fail to take into account semantic similarities among scales. Thus, a gap exists which concerns the objective evaluation of scale language.

To tackle this important gap, we propose in this paper the development of a validity measure – termed *manifest validity* – to test for obvious language-driven survey results. Studying the semantic similarities between survey items, researchers can extract the semantic difference between different scales and examine whether the respondents employed deep or shallow processing during questionnaire analysis, thereby essentially understanding whether language based instinctual responses were triggered in the respondents. This proposed validity

measure has the potential to be a required ingredient in most future questionnaire based studies, opening up new areas of inquiry within psychometrics, with the potential to improve questionnaire design and provide a better understanding of measured constructs.

The paper continues as follows. First, we briefly review some foundations of survey design following with an introduction to the proposed approach – latent semantics analysis (LSA). LSA is a method for extracting and representing the contextual-usage meaning of words and sentences. LSA functions analogously to a human brain in terms of language use and understanding when providing "interpretations" of text [3]. Thus, LSA is capable of computing the semantic similarities or distances of language statements. We then demonstrate the usability and potential value of this approach by applying it to data from previously published surveys, demonstrating that a new validity measure is needed. We conclude with a discussion of the insights obtained from these preliminary findings and of the potential contributions and future applications of the proposed approach.

2. Survey Research

Literally thousands of academic theories rely on survey research. Such surveys have been used in a multitude of theories to predict behaviors as diverse as use of oral rehydration therapy to web design practices, as well as hundreds of different types of human behavior. For example, the initial papers developing three popular theories *Theory of Reasoned Action* [5], *Theory of Planned Behavior* [6], and the *Technology Acceptance Model* [7, 8] alone have received close to 10,000 citations during the last 25 years.

Survey research entails the development of questionnaires often measuring latent constructs such as attitudes and perceptions. Constructs are measured using measurement scales –collections of language based statements (termed items) to which respondents express their level of (dis)agreement. Two main problems are involved with the use of such scales: the quality of the responses and the quality of the questions. First, because of the popularity of surveys, the respondent population is often asked to fill out numerous questionnaires, leading to questionnaire fatigue and reluctance to respond conscientiously. Apart from the risk of non-response bias, subjects' honesty, memory, and motivation will also affect the quality of the results. In some cases, respondents may have forgotten their reasons for conducting a certain action, or may not be willing to think deeply about answers to questions. In such cases, respondents are likely to revert to knowledge built into language itself (i.e. respondents will simply reflect the written language of the sentence)

A second problem concerns the quality of the scales. Serious unanswered questions still remain about *what* questionnaire research actually measures. This problem falls under the domain of survey validity and is the focus of this paper. Specifically, we study an under-explored area of questionnaire development – *the choice of language combinations in scales items*. We propose that the underlying semantic structure of scale items and the specific choices of words impacts survey responses by tapping into features of language itself rather than real-life settings. Before discussing further the shortcomings of current validity measures we first provide a short introduction to survey reliability and validity measures (based on Pedhazur & Schmelkin, [9] and Singleton & Straits, [10]).

2.1. Survey validity and reliability

In constructing measurements for research variables, the reliability and validity of the constructed scale should be assessed. Reliability is mainly concerned with the consistency of the scale: do repeated measures yield similar results? While validity is concerned with the question of whether the scale indeed measures what it purports to measure. Using a target metaphor, reliability measures whether all hits cluster closely around the same location, while validity measures whether the cluster is indeed at the center of the target.

The simplest test of reliability is the test-retest procedure, in which the same unit is measured twice at different occasions, and results are then correlated. More robust measures focus on the internal consistency of the full scale, evaluating the homogeneity in the individual items. A commonly used measure of internal consistency is Cronbach's alpha, measuring the extent to which all individual items correlate with each other.

Unlike reliability, validity cannot be assessed directly, as we have no knowledge of the true value of the construct. Rather, we use several forms of validation to ensure high internal validity of the constructed scale. First, subjective measures such as content validity and face validity can be applied to assess the extent of domain coverage and comprehensibility of the survey, respectively. Second, construct validity is used to examine which construct the scale actually measures. The measurement of construct validity consists of evaluating: convergent validity (the extent to which the scale correlates with other measures of the same construct); discriminant validity (the extent to which the scale does not correlate with other constructs from which it is supposed to differ); and nomological validity (the degree to which the construct predicts other constructs as stated by the theory used).

2.2. Shortcomings of existing measures

The above measures of reliability and construct validity are mechanical in nature and rely heavily on numerical outcomes, such as rating of items. At the same time, the measures of face and content validity are highly subjective. Thus, while some objective validity measures exist they lack the semantic focus; and while some measures focus on the semantics of the survey, they are not objective. The outcome is an important deficiency of current validity and reliability measures. In particular, current measures are incapable of identifying important semantic phenomena such as distinguishing between deep vs. shallow language processing of scales items or knowing when the wording of the items elicits instinctual responses. Referring to the former, shallow vs. deep processing is often seen as a continuum where shallow processing reflects the memorization of text and use of surface features of the language during processing of text, while deep processing involves reflecting on the nuances and meaning of text and attaching personal relevance to the information provided [11,12]. Survey respondents may shallowly process questionnaire items for numerous reasons such as limited involvement in the survey topic, limited knowledge of the survey's context, lack of experience, or faulty questionnaire design. Experiments have shown that many readers tend toward processing text in incomplete or shallow manners, and that less-skilled readers often fail to understand global context of a sentence before reacting to it [11, 12]. In other words, what level of understanding do the respondents have of the questions they are rating? The extent of deep understanding of the context of the study (either through extensive experience or through study) is one important enabler of deep processing. Unfortunately, simplistic methods of face validity rely on the subjective evaluation of field experts and potential respondents and are incapable of accurately assessing comprehension

Moving to the latter, because language and environment have co-evolved for millennia [2], it is likely that some language cannot be properly distinguished from experiential knowledge in that it elicits instinctual responses. Of course, if the goal is to measure language's ability to retain knowledge, this is not a problem. However, if the goal of a study is to measure personally experienced knowledge or attitudes, such language effects may represent a vexing problem. Again, current measures might be limited in identifying such problems in questionnaire scales.

To overcome this deficiency of current measures we propose the development of a new measure of validity, namely manifest validity. Manifest validity will offer an objective measure of questionnaire language and connotations that is currently lacking from validity measures (objective in the sense that it will not rely on the language perception of the individual researcher). Specifically, manifest validity is expected to support researchers during the data analysis stage in that researchers can compare measures of manifest validity (evaluating the extent of semantic difference between different scales) to item correlations computed from actual responses. In cases where there is little difference between distances proposed by the semantic analysis of items and distances proposed by correlation coefficients, the respondents are more likely to have employed shallow processing during questionnaire analysis. To the extent that such semantically driven results are the result of the language of the questionnaire itself rather than the questionnaire fatigue of a specific response population, standard questionnaires may need redevelopment.

We turn next to describe how manifest validity can be developed using Latent Semantic Analysis [13] – a method that enables us to measure the semantic similarity among text items.

3. Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a theoretically-based method for extracting and representing the contextual-usage meaning of words, using statistical computations. The underlying idea of LSA is that the aggregate of all the word contexts in which a given word does and does not appear provides a set of constraints that determines the similarity of meaning of words, and sets of words, to each other [14]. 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. This representation can be used to compare similarities in different documents (collections of words).

To apply LSA to some domain of inquiry, a large set of documents is represented in a sparse matrix of documents (columns) vs. words in those documents (rows), as shown in the example in Tables 1 and 2. Each cell in the matrix in Table 2 represents the number of times that the row's word appears in the column's document. Note that for this example focus on only a subset of words that are interesting for this topic (here denoted in bold in Table 1)[2].

[2] Note that the typical document collection LSA starts with will be between 10,000 and 100,000s documents

Document	Text
D1.	The **IT** Factor
D2.	**IT** and **IS**, a Respecification of **Theory**
D3.	Only **TAM** Matters in **IS** and **MIS**
D4.	Future Perfect, **TAM** uber alles
D5.	**Design Theory**: The New Pink
D6.	**Mathematical** Understanding in **Operations Research**
D7.	**Operations Research Design** and **Mathematics**

Table 1: LSA example text (documents)

	D1	D2	D3	D4	D5	D6	D7
IT	1	1	0	0	0	0	0
IS	0	1	1	0	0	0	0
TAM	0	0	1	1	0	0	0
MIS	0	0	1	0	0	0	0
Theory	0	1	0	0	1	0	0
Design	0	0	0	0	1	0	1
Mathematics	0	0	0	0	0	1	1
Operations	0	0	0	0	0	1	1
Research	0	0	0	0	0	1	1

Table 2: Term-document matrix

The matrix in Table 2 is then normalized and weighted and submitted to a Singular Value Decomposition (SVD) (for an in-depth explanation of SVD, the reader is referred to Martin and Berry [15] and Berry, Drmac, and Jessup [16]). The outcome is generally referred to as a *semantic space* with k dimensions. While there is clearly structure to each dimension, LSA does not depend on an interpretation of each dimension, but rather examines a text unit's structure across all dimensions.

The SVD algorithm approximates the above matrix in lower dimensionality (the matrix in Table 2 could be perfectly replicated at seven dimensions – the smallest of terms and documents). For LSA to work, it is critical to reduce dimensionality, a process that facilitates LSA's ability to generalize from a relatively small sample to most texts not yet encountered. For display ease, we keep the number of dimensions at two as illustrated in Figure 1. In Figure 1 two sets of data points are represented: the squares represent the original documents (D1 through D7) while the diamonds represent the individual terms (the first column in Table 2).

Each word and each document shown in Table 2 is now located in a two-dimensional vector space, which represents a *semantic space* – a representation of the distances between text units – but one that in this case will be very limited due to the small sample and low dimensionality.

Notice that at this dimensionality, the words *mathematics*, *operations*, and *research* are considered identical (they are all represented by a single point), but additional information from more dimensions would likely show that they are somewhat different. It is worth noting that even for this tiny example, when asked, most experts in the IS and Operations field identify documents one through four as being similar in that they are all related to the IS field, and documents six and seven being from the operations area of research. Indeed as shown in Figure 1, D1 through D4 cluster around the lower left corner of the space while D6 and D7 are both at the upper right corner. Document five is generally considered different from both groups, but also somewhat related to each, located in the space between the two clusters. It is especially noteworthy that the similarity measure between MIS and IT suggest that they are virtually indistinguishable in spite of never occurring in the same document. Herein lies the strength of LSA; the context of every word in the whole document set is taken into consideration during the solving of a large set of simultaneous equations in the SVD.

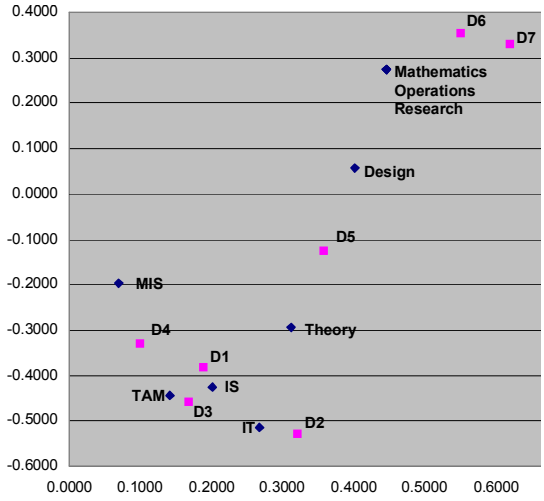

Figure 1: A two-dimensional semantic space example

With the semantic space in place, new texts such as questionnaire statements may now be "projected" into this space (in the same way that the original documents were introduced), and distances between them may be calculated. The main drawback from the approach is that only words known to the semantic space – the nine words in Table 2, in this particular example – can be used for future analysis. This, however, tends not to be a problem when a large and representative sample of texts is selected for the initial creation of the semantic space. Finally, most commonly, the cosine measure rather than Euclidian distance measure between items is considered the most valid measure.

3.1. LSA and Manifest Validity

Since introduced by Deerwester et al. [13], LSA has been used to retrieve documents [e.g., 17], automatically summarize text [e.g., 18] and video [e.g., 19], grade essays [e.g., 20], automatically construct thesauri [e.g., 21], translate between languages [e.g., 22], as well as to

cluster and extract knowledge from text [e.g., 23]. Of these many applications, the ability to detect semantic similarities among text items is most vital to our study.

We propose that LSA can be effectively applied to scale items to predict their semantic closeness. Specifically by constructing a relevant semantic space (i.e. for the Information Systems research domain) we can test scale items to identify the semantic distance between them. Semantically closer items measure similar things, while semantically distant items measure different semantic things. Consider, for example, Gefen et al.'s [24] *ease of use* scale item "It is easy to become skillful at using the Web site". Using LSA we can characterize this item as highly semantically similar to their *perceived usefulness* scale item: "The web site is useful for searching and buying CDs/books" (cosine: 0.74; correlation: 0.65). And indeed, the latter item was dropped by the authors due to "high degree of residual variance with other items" [24, p. 68]. However, a second *perceived usefulness* item: "The Web site improves my performance in CD/book searching and buying" was not dropped (correlation: 0.58), even though it still shows a high semantic relationship to the same *ease of use* item (cosine: 0.57; While we cannot yet propose cutoff points for cosines between items from different scales, both .74 and .57 are rather high). It is not our intention to critique Gefen et al's highly cited study which was published in a premiere IS journal, but rather to show that sometimes semantically similar items may be different enough to be distinguishable by principal component analysis, yet similar enough to ensure relationships between different scales. It is also not meant to suggest that LSA cosines will mirror actual sizes of correlations (as was the case here), but rather that sometimes the results of questionnaire research will mirror knowledge built into our shared language rather than anything specific to the setting of the study, in this case attitudes towards online shopping.

By now we have identified the gap in current validity measures and highlighted the potential ability of LSA to close this gap. To further strengthen our argument we provide an example in which we predict the outcomes of scale items correlations in 22 Information Systems (IS) studies. What we will argue is that if the cosines derived from LSA can significantly and highly predict *actual* inter-item correlations then there is some inherent semantic problem in the scale. Essentially, in such cases survey outcomes are highly predicted by an *automated* semantics based application (LSA) without any human perceptions involved. Hence, the purpose of this example is to show that the measures derived from LSA are indeed good potential measures of the semantic validity of scales, at least for one example of behavioral studies.

4. Applicability of LSA to Predict Survey Outcomes

The specific steps involved in applying LSA to survey data involve: (1) creation of a semantic space; (2) computation of cosines as measures of the semantic similarities among items; and (3) evaluation of the extent to which these cosines predict item correlations, as computed from the survey data. Below we describe each of these steps in more detail.

4.1. Creation of the semantic space

A semantic space may be developed as described in the earlier example, except with at least 30,000 carefully selected documents to enable the semantic space an appropriate "understanding." The documents –often paragraphs from larger documents such as books and articles– appropriate for a semantic space would typically mirror the general knowledge in the domain of interest. For example, a general purpose semantic space (aka the TASA space) which mirrors the education received by an individual through high school is available through lsa.colorado.edu[3]. Unfortunately, due to its age and focus this semantic space is not appropriate for our purpose, as it does not contain documents that would mirror an understanding of neither information systems nor general social science research.

One approach would thus be to create a new semantic space including tens of thousands of newspaper and magazine articles with sufficient general- as well as IS-related knowledge that will provide a good understanding of IS word usage. However, by building a semantic space specialized for the task at hand, we would be open for criticism that we (over)-fit a solution to our problem. Instead, we decided to go with an existing semantic space that we had no hand in creating, and that is available for reviewers and readers in general to replicate the results of this study. The semantic spaces on the lsa.colorado.edu website provide such a setting, and we therefore decided to use a semantic space available from that website.

Not developed by the research team, the Computer Supported Collaborative Learning (CSCL) space was constructed based on a sample of documents from the TASA set as well as a sample of articles related to computer supported collaborative learning. While the latter set of documents would provide a rough understanding of many concepts and words used in IS studies, none of these documents have publication dates after 2001, and the majority of documents are published around 1997. It was clear that the CSCL space was far

[3] The lsa.colorado.edu Web site is provided by a research group unaffiliated with the authors of this paper, and the TASA semantic space is listed as "General_Reading_up_to_1st_year_college."

from ideal for our purposes – in fact, any results found using this space should only be seen as a starting point for later analysis. In spite of these shortcomings, it was the most appropriate *publicly available* semantic space, and all the analysis in this paper is done with this space, including the short example from Gefen et al. [24] outlined above. Based on conversations with the Colorado researchers running the lsa.colorado.edu website, the space was created from 33,419 documents with 62,434 unique words. Our later analysis of the documents that went into the space suggests that it contained readings from among others, *A People's History of the United States, the Cuban Missile Crisis*, a book entitled *Education and Mind*, as well as paragraphs from the 1995, 1997, and 1999 CSCL conference held in Boulder, CO.

4.2. Computation of cosines

We were able to find 22 papers for our analysis. The criterion for paper selection was the availability of an inter-item correlations table and of the full questionnaire. Hence, each of the papers selected included both the detailed language of questionnaire items as well as an inter-item correlation table based on the data collected in the study. The final list of papers selected is presented in the first column of Table 3. To find suitable papers for this analysis we examined ten years of publications from four IS journals namely: Information Systems Research, MIS Quarterly, Journal of Management Information Systems, and Information and Management, as well as contacted authors directly[4]. The final set of papers analyzed included one unpublished study, three studies published earlier or during the time of the documents used to create the semantic space, and 18 studies published after the same period.

The scale items of each selected paper (e.g. "The Web Site is easy to use", "It is easy to become skillful at using the Web Site", etc.) were projected onto the CSCL semantic space, in order to calculate the semantic distances –or cosines- between them. Cosines were calculated for every pair of scale items so that an inter-item cosine table was created. At the final stage of our analysis we compared our calculated cosines to the correlations reported in the paper and computed the R^2 statistic based on this comparison.

4.3. Predictive power of cosines

The final step in our analysis included examining the calculated cosines vis-à-vis the actual inter-item correlations obtain from the survey responses. Specifically, we employed a simple regression analysis, regressing the cosines obtained from the semantic space with the correlations obtain in the original studies. The R^2's are reported in Table 3.

As may be seen in Table 3, the correlations found in these existing studies could be predicted with an R^2 between .00 and .63. The average R^2 was .22. In other words – there is clear evidence of a relationship between the semantics of many questionnaire items and respondents' perceptions of the real-life meaning of these items. The magnitude of this relationship testifies to the semantic validity of a study, with lower R squares indicating potentially higher semantic validity and vice-versa. Thus for those cases where the R square scores are high, our analysis indicates a potential semantic problem stemming from the fact that actual perceptions of respondents are well predicted by the wording of the items. It is especially interesting to see that among the seven studies that LSA was able to best predict, the highest four were all in the educational arena, an area identified by Umbach [1] as especially vulnerable to survey fatigue. Of the remaining three studies, two were of CEO's, CIO's and senior managers, all high targets for survey researchers, and likely to have little discretionary time.

Overall, these results are encouraging in that they reveal the ability of LSA to detect underlying semantic problems, even when the semantic space used was not specifically made for the academic field of these papers and otherwise not created with a very large set of documents (36,000 document in this case). Thus, we expect that an appropriately crafted semantic space will yield even better results.

[4] This of course is not intended as an extensive review of the literature but rather to facilitate and example demonstrating the applicability of LSA in measuring the semantic validity of a questionnaire. We would also like to thank the authors who agreed to share this information with us.

Paper	Subjects	Predictive ability (R^2)	Comments
[25]	86 consumer electronics firms (46%-67% response rate)	**.00**	Inappropriate study. Just one scale and reference to real-world events rather than attitudes
[26]	150 managing directors of South African companies (19% response rate).	**.00**	
[27]	342 knowledge workers from 25 Taiwanese companies (68% response rate)	**.07***	Taiwanese respondents and language
[28]	284 IS managers responsible for first implementing corporate website (19% response rate)	**.08*****	
[29]	251 CIOs and VP's or Directors of MIS (48% response rate)	**.09*****	
[30]	140 respondents from six international organizations (75.3% response rate)	**.09*****	Seemingly Asian companies. Language unknown.
[31]	173 members of a national legal professional association in the United States (29% response rate)	**.11***	Correlation matrix provided by authors.
[32]	365 firms (40% response rate)	**.14*****	
[33]	421 specialist physicians in Hong Kong (24% response rate)	**.15*****	Chinese language
[34]	116 respondents from five international organizations	**.15*****	
[35]	565 Project Manager IS SIG members of Project Management Institute (32% response rate)	**.18*****	
[36]	122 online customers (response rate 12%)	**.19*****	
[37]	87 IS and logistics managers (21% response rate)	**.19*****	
[38]	629 Korean managers	**.20*****	Korean language
[39]	120 independent agencies	**.24*****	
[40]	members of 24 Undergraduate student teams (competition and rewards)	**.32*****	
[41]	523 members of family panel (35% response rate)	**.34****	
[42]	161 CIOs (13% response rate)	**.36*****	
[43]	226 faculty members	**.43*****	
[44]	77 CEOs and 166 senior managers	**.45*****	
[45]	274 university employee users of information system (30% response rate)	**.58*****	
[24]	213 graduate and undergraduate students	**.63*****	

Table 3: Examination of existing studies *** significant at $p < .001$, ** significant at $p < .01$, * significant at $p < .05$.

5. Discussion

In this paper we introduced the idea of manifest validity to study language based biases in scale items and questionnaire design. We argued that there is currently a gap in our methods of evaluating questionnaire validity, with current methods incapable of accurately detecting when respondents employ shallow processing. Two important propositions can be made based on our analysis so far. First, since high cosines between any two text-items represent semantic similarities we propose that high cosines are undesirable for items representing different constructs. Moreover, we propose that LSA can be extremely useful in developing survey items by selecting items which are semantically similar to the definition of the construct which they purport to measure. Thus, LSA can add an important semantic dimension to existing measures of construct and discriminant validities.

Second, using the example of 22 papers we demonstrated that cosine predictions are significant while vary in magnitude. We propose that the existence of high R^2 values between inter-item cosines and correlations is indicative of an underlying semantic problem that should be further investigated. What this result means is that the semantic distance between scale items (as reflected in the calculated cosines) can provide a good estimate of underlying semantic problems, such as the extent to which respondents employed shallow processing when filling out a survey. Future research should study the link between cosines and shallow processing in more depth at both the survey as well as the individual respondents level. For example, studies may focus on detecting *which* individual respondents used shallow processing, thereby solving a major problem in survey research. Furthermore, future research can extend the study of shallow processing beyond the survey method and into other research designs.

The example presented in the previous section also demonstrates the potential usefulness of LSA as the method for testing semantic aspects of questionnaire scales. The high correlations imply that the cosine measure can be used to determine scales correlation *a priori*, as well as provide an alternative to the subjective face validity measure. An interesting finding from this example is that Lee et al.'s paper used a set of English-text items translated into Korean. Comparing the cosines between the original English-text items and the correlations from the Korean-text responses still yielded a relationship, suggesting that this approach can work well even before foreign-language semantic spaces are created.

In the introduction to this paper we discussed the problem of respondents fatigue and how it affects the outcome of surveys, both in terms of low response rates and in terms of the quality of responses. Indeed, we showed how – likely because of shallow processing by respondents – LSA cosines often behave in similar ways as correlations resulting from the analysis of large respondent datasets.

Finally, a methodological insight from this preliminary example is that using LSA, the constructed semantic space provides the domain of knowledge upon which language choices are evaluated. Therefore, the creation of a wide semantic space is crucial for the development of manifest validity. Moreover, we call for the inclusion of full inter-item correlation tables in published papers to enable their semantic evaluation, a practice that is not yet common within IS publications.

One potential limitation of this paper is the use of a limited semantic space and a relatively small number of papers. However, since this paper is intended to introduce new thinking on survey validity and mainly demonstrate the applicability of such approach to social science research in general, and IS research in particular, we argue that this limitation does not lessen the contribution of the paper. Our ability to show, even on such small scale, the applicability of LSA to analyze survey items and the ability of the calculated cosines to predict actual correlations among scale items demonstrates the potential significance of this work. We further note that this study focuses on reflective rather than formative items. Obviously items of formative constructs are expected to be semantically distant from each other, and we did not study such items in this work,

A direct continuation of this study will develop a more holistic application of manifest validity, to provide additional measures of construct and content validity. First, more exploration is needed on the specific level that this validity measure can be applied at. Our study drew conclusions at the questionnaire level, however analysis can also involve pair-wise comparisons of constructs (for a greater focus on discriminant validity) for example. In addition specific guidelines concerning acceptable threshold and applied measures of semantic validity still need to be developed. Second, an important aspect of questionnaire content not tested by current validity measures is the semantic distance between different scales. Only when scales are found to be different through both factor analysis and semantic analysis are findings likely to be truly meaningful. We intend to extend manifest validity to also provide an additional measure of construct validity, discriminating between constructs according to their semantic differences. Third, some studies often use open-ended questions as well. Distance between items of a study and individual open-ended question responses may also be measured to categorize the text in terms of existing knowledge,

or determine that the open-ended question is related to another topic not covered by the questionnaire. The latter is a step towards formalizing a measure of content validity. Finally, we also intend to enhance the application of manifest validity to survey creation by creating a search engine in an unconventional sense. By superimposing a database of published questionnaire items on a semantic space, researchers may insert their questionnaire items and get feedback on similar items in the semantic space, thereby saving much time on development of new questionnaire items, and ensuring good incremental science.

5. Conclusion

This study promises to open up a new area of inquiry within psychometrics, and has the potential to improve questionnaire design as well as to provide a better understanding of *what* questionnaire scales items measure. Researchers can use manifest validity to develop semantically stronger scales, as well as to understand whether survey respondents employed deep vs. shallow processing of items. The latter can be supported by identifying a *minimum* semantic distance between items not belonging to the same scale. Using this distance, researchers can be relatively certain that they are not overtly tapping language rather than context and experience. Moreover, manifest validity can also support new and innovative research by alerting researchers of their use of non-traditional language combinations and recommending the use of subjective measures of face and content validation in tandem with the manifest validity.

6. References

[1] P. D. Umbach, "Getting Back to the Basics of Survey Research," *New Directions for Institutional Research*, vol. 127, pp. 91-100, 2005.

[2] M. D. Hauser, N. Chomsky, and W. T. Fitch, "The Faculty of Language: What is it, Who Has it, and How Did it Evolve?," *Science*, vol. 298, pp. 1569-1579, 2002.

[3] T. K. Landauer, "LSA As a Theory of Meaning," in *Handbook of Latent Semantic Analysis*, D. S. M. Thomas K Landauer, Simon Dennis, and Walter Kintsch Ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2007.

[4] T. K. Landauer, "On the Computational Basis of Cognition: Arguments from LSA," in *The Psychology of Learning and Motivation*, B. H. Ross, Ed. New York, NY: Academic Press, 2002.

[5] I. Ajzen and M. Fishbein, *Understanding Attitudes and Preicting Social Behavior*. Englewood Cliffs, NJ: Prentice-Hall, 1980.

[6] I. Ajzen, "From intentions to actions: A theory of planned behavior," in *Springer series in social psychology* J. Kuhl and J. Beckmann, Eds. Berlin: Springer, 1985, pp. 11-39.

[7] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly*, vol. 13, pp. 319-340, 1989.

[8] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw, "User Acceptance of Computer Technology: A Comparison of Two Theoretical Models," *Management Science*, vol. 35, pp. 982-1003, 1989.

[9] E. J. Pedhazur and L. P. Schmelkin, *Measurement, Design, and Analysis: An Integrated Approach*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1991.

[10] R. A. Singleton and B. C. Straits, *Approaches to Social Research*, 3rd ed. New York, NY: Oxford University Press, 1999.

[11] A. J. S. Sanford, J. Molle, and C. Emmott, "Shallow Processing and Attention Capture in Written and Spoken Discourse," *Discource Processes*, vol. 42, pp. 109-130, 2006.

[12] R. R. Schmeck, F. Ribich, and Ramanaiah, "Development of a Self-Report Inventory for Assessing Individual Differences in Learning Processes." *Applied Psychological Measurement,* vol. 1, pp. 413-431, 1977.

[13] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, vol. 41, 1990.

[14] T. K. Landauer, P. W. Foltz, and D. Laham, "Introduction to Latent Semantic Analysis," *Discourse Processes*, vol. 25, pp. 259-284, 1998.

[15] D. I. Martin and M. W. Berry, "Mathematical Foundations Behind Latent Semantic Analysis," in *Handbook of Latent Semantic Analysis*, D. S. M. Thomas K Landauer, Simon Dennis, and Walter Kintsch Ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2007.

[16] M. W. Berry, Z. Drmac, and E. R. Jessup, "Matrices, Vector Spaces, and Information Retrieval," *SIAM Review*, vol. 41, pp. 335-362, 1999.

[17] A. Kontostathis and W. M. Pottenger, "A framework for understanding Latent Semantic Indexing (LSI) performance," *Information Processing & Management*, vol. 42, pp. 56-73, 2006.

[18] S. T. Yuan and J. Sun, "Ontology-based structured cosine similarity in document summarization: With applications to mobile audio-based knowledge management," *Ieee Transactions on Systems Man and Cybernetics Part B-Cybernetics*, vol. 35, pp. 1028-1040, 2005.

[19] Y. H. Gong and X. Liu, "Video summarization and retrieval using singular value decomposition," *Multimedia Systems*, vol. 9, pp. 157-168, 2003.

[20] M. A. Hearst, "The debate on automated essay grading," *Ieee Intelligent Systems & Their Applications*, vol. 15, pp. 22-27, 2000.

[21] M. Hagiwara, Y. Ogawa, and K. Toyama, "PLSI utilization for automatic thesaurus construction," in *Natural Language Processing - Ijcnlp 2005, Proceedings*, vol. 3651, *Lecture Notes in Artificial Intelligence*, 2005, pp. 334-345.

[22] A. Fujii and T. Ishikawa, "Japanese/English cross-language information retrieval: Exploration of query translation and transliteration," *Computers and the Humanities*, vol. 35, pp. 389-420, 2001.

[23] K. R. Larsen and D. E. Monarchi, "A Mathematical Approach to Categorization and Labeling of Qualitative Data: the Latent Categorization Method," *Sociological Methodology*, vol. 34, pp. 349-392, 2004.

[24] D. Gefen, E. Karahanna, and D. W. Straub, "Trust and TAM in Online Shopping: An Integrated Model," *MIS Quarterly*, vol. 27, pp. 51-90, 2003.

[25] F. Kaefer and E. Bendoly, "Measuring the Impact of Organizational Constraints on the Success of Business-to-business E-commerce Efforts: A Transactional Focus," *Information & Management*, vol. 41, pp. 529-541, 2004.

[26] A. Molla and P. S. Licker, "eCommerce adoption in developing countries: a model and instrument " *Information & Management*, vol. 42, pp. 877-899, 2005.

[27] R. McHaney, R. Hightower, and J. Pearson, "A Validation of the End-user Computing Satisfaction Instrument in Taiwan," *Information & Management*, vol. 39, pp. 503-511, 2002.

[28] R. C. Beatty, J. P. Shim, and M. C. Jones, "Factors Influencing Corporate Web Site Adoption: A Time-based Assessment," *Information & Management*, vol. 38, pp. 337-354, 2001.

[29] A. H. Segars and V. Grover, "Strategic Information Systems Planning Success: An Investigation of the Construct and Its Measurement," *MIS Quarterly*, vol. 22, pp. 139-163, 1998.

[30] C.-S. Ong, J.-Y. Li, and Y.-S. Wang, "Factors affecting engineers' acceptance of asynchronous e-learning systems in high-tech companies," *Information & Management*, vol. 41, pp. 795-804, 2004.

[31] M. M. Wasko and S. Faraj, "Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice," *MIS Quarterly*, vol. 29, pp. 35-57, 2005.

[32] H. Tanriverdi, "Performance Effects of Information Technology Synergies in Multibusiness Firms," *MIS Quarterly*, vol. 30, pp. 57-77, 2006.

[33] P. J. Hu, P. Y. K. Chau, O. R. L. Sheng, and K. Y. Tam, "Examining the Technology Acceptance Model Using Physician Acceptance of Telemedicine Technology," *Journal of Management Information Systems*, vol. 16, pp. 91-112, 1999.

[34] Y.-S. Wang, "Assessment of Learning Satisfaction with Asynchronous Electronic Learning Systems," *Information & Management*, vol. 41, pp. 75-86, 2003.

[35] W. Xia and G. Lee, "Complexity of Information Systems Projects: Conceptualization and measurement Development," *Journal of Management Information Systems*, vol. 22, pp. 45-83, 2005.

[36] A. Bhattacherjee, "Individual Trust in Online Firms: Scale Development and Initial Test," *Journal of Management Information Systems*, vol. 19, pp. 211-241, 2002.

[37] K. S. Soliman and B. D. Janz, "An Exploratory Study to Identify the Critical Factors Affecting the Decision to Establish Internet-based Interorganizational Information Systems," *Information & Management*, vol. 41, pp. 697-706, 2004.

[38] Y. Lee, K. A. Kozar, and K. R. Larsen, "Avatar E-mail Use: A Theoretical Integration and Extension," under review.

[39] A. Zaheer and N. Venkatraman, "Determinants of Electronic Integration in the Insurance Industry: An Empirical Test," *Management Science*, vol. 40, pp. 549-566, 1994.

[40] R. C.-W. Kwok, J. Ma, and D. R. Vogel, "Effects of Group Support Systems and Content Facilitation on Knowledge Acquisition," *Journal of Management Information Systems*, vol. 19, pp. 185-229, 2002.

[41] N. F. Awad and M. S. Krishnan, "The Personalization Privacy paradox: An Empirical Evaluation of Information Transparency and the Willingness to be Profiled Online for Personalization," *MIS Quarterly*, vol. 30, pp. 13-28, 2006.

[42] G. S. Kearns and A. L. Lederer, "The Impact of Industry Contextual Factors on IT focus and the Use of IT for Competitive Advantage," *Information & Management*, vol. 41, pp. 899-919, 2004.

[43] W. Lewis, R. Agarwal, and V. Sambamurthy, "Sources of Influence on Beliefs about Information Technology Use: An Empirical Study of Knowledge Workers," *MIS Quarterly*, vol. 27, pp. 657-678, 2003.

[44] J. Karimi, T. M. Somers, and Y. P. Gupta, "Impact of Environmental Uncertainty and Task Characteristics on User Satisfaction with Data," *Information Systems Research*, vol. 15, pp. 175-193, 2004.

[45] A. Rai, S. S. Lang, and R. B. Welker, "Assessing the Validity of IS Success Models: An Empirical Test and Theoretical Analysis," *Information Systems Research*, vol. 13, pp. 50-69, 2002.