Abstract

Psychological measurement and theory are afflicted with an ongoing proliferation of new constructs and scales. Given the often redundant nature of new scales, psychological science is struggling with arbitrary measurement, construct dilution, and disconnection between research groups. To address these issues, we introduce an easy-to-use online application: the Semantic Scale Network. The purpose of this application is to automatically detect semantic overlap between scales through Latent Semantic Analysis. Authors and reviewers can enter the items of a new scale into the application, and receive quantifications of semantic overlap with related scales in the application's corpus. Contrary to traditional assessments of scale overlap, the application can support expert judgements on scale redundancy without access to empirical data or awareness of every potentially related scale. After a brief introduction to measures of semantic similarity in texts, we introduce the Semantic Scale Network and provide best practices for interpreting its outputs.

Keywords: Scale development; Scale proliferation; Network analysis; Research infrastructure; Latent semantic analysis; Decision support system

Supplementary materials: The application can be found on rosenbusch.shinyapps.io/semantic_net. Data and source code for the application can be found on https://osf.io/y87pe. A backup server is provided by the University of Tilburg (see description on OSF).

The Semantic Scale Network: An online tool to detect semantic overlap of psychological scales and prevent scale redundancies

Psychologists often rely on scales to measure psychological constructs, such as attitudes (Bar-Anan & Vianello, 2018), traits (Simms, Zelazny, Williams, & Bernstein, 2019), emotions (Pekrun, Vogl, Muis, & Sinatra, 2017), and beliefs (Muis, Duffy, Trevors, Ranellucci, & Foy, 2014). These scales usually consist of a set of questions or statements that participants respond to by indicating their approval or agreement (Loewenthal & Lewis, 2001). Often, researchers create new scales, which can run the risk of being redundant with existing scales (Bruner, 2003; Haynes & Lench 2003; Nimon, Shuck, & Zigarmi, 2016; Shaffer, deGeest & Li, 2016). Whereas some researchers actively investigate and combat scale (and construct) redundancies in their fields of expertise (e.g., Banks, McCauley, Gardner, & Guler, 2016; Cole, Walter, Bedeian, & O'Boyle, 2012; Morrow, 1983; Reeve & Basalik, 2014; Roodt, 2004), initial publications of new scales often do not sufficiently justify their incremental value (Haynes & Lench 2003; Hunsley & Meyer, 2003; Sechrest, 1963). Yarkoni (2010) showed that genetic algorithms can condense 203 psychological scales into 181 items, which can through recombination accurately capture the variance of the original scales (an alternative algorithm to condense multi-facet scales is discussed by Olaru, Schroeders, Hartung, & Wilhelm, in press). While the author introduced the method as a way to abbreviate scales, we believe it also speaks to the large amount of overlap found in psychological scales. Such overlap and redundancies can only be expected to increase in the future as the mass of published psychological scales keeps growing.

Problems resulting from this ongoing proliferation of scales are manifold. First, researchers have to decide which scale to use for measuring constructs, which becomes increasingly difficult if many alternative scales have been published (Terwee et al., 2007).

Second, the content of psychological constructs cannot be expected to be completely stable across scales. Thus, incompatible research findings can emerge, leading to separated research strings and diluted construct interpretations (Cole et al., 2012). Third, with the growing mass of scales, the chances of finding spurious correlations between constructs is relatively high. Not only do alternative scales inflate the danger of Type I errors, as interchangeable scales can lead to more tests per study, but they also increase the likelihood that a pair of scales for the respective constructs have similar item phrasings, which can induce spurious correlations (Arnulf, Larsen, Martinsen, & Bong, 2014; Arnulf, Larsen, Martinsen, & Egeland, 2018; Clark & Watson, 1995; Gefen & Larsen, 2017; MacKenzie, Podsakoff, & Podsakoff, 2011; Maul, 2017). Such spurious correlations can easily be misinterpreted as convergent validity whereas they actually, given the linguistic overlap between scales, indicate measurement reliability (Campbell, & Fiske, 1959). In short, redundant scales threaten some of the basic requirements of psychological science such as standardized measurement and well-understood constructs. In order to maintain and improve the quality of scale-driven research, the proliferation of unneeded scales needs to be prevented.

In the next sections, we review current strategies of assessing scale redundancies. Importantly, we highlight how automatic analyses of semantic overlap between scales can complement current methods. Subsequently, we demonstrate how such semantic similarity between questionnaire texts can be quantified. Finally, we introduce a "shiny" application (R-based web application; Chang, Cheng, Allaire, Xie, & McPherson, 2018) that automatically assesses semantic overlap between new scales and previously published scales: the Semantic Scale Network.

Beyond Correlational Analyses and Expert Judgement

To date, the predominant approach to identify redundancy between scales has been to correlate participant scores on different candidate scales, with high correlations indicating potential redundancy (e.g., Cole et al., 2012; Le, Schmidt, Harter, & Lauver, 2010). However, there are multiple problems inherent to this approach. First, researchers have to decide *before* data collection which scales might be redundant to the new instrument, as they need participant data to quantify shared variance. Despite best efforts to stay up to date, it is difficult to be aware of every scale that might be relevant for one's research. Relevant scales are often published under different names, or even in different disciplines, and might therefore escape researchers' attention. Second, researchers need to collect data for all scales from the same test subjects, which might be problematic if there are too many related scales for each participant to fill out. Third, there is no cut-off between high convergent validity and redundancy (e.g., Cole et al., 2012), mostly because a generic cut-off value cannot do universal justice to assessing redundancy (cf. discussion of reliability coefficients by Lance, Butts, & Michels, 2006). Instead of exclusively relying on numerical analyses, it is therefore reasonable to assess redundancy through theory-guided justifications of a scale's incremental value based on item content. Accordingly, an expert analysis of item content across scales is necessary beyond the computation of a correlation score. Fourth, usually only the publisher of a new scale conducts and reports empirical tests of the new scale. Reviewers typically do not have the means to collect data and run additional analyses to assess the uniqueness of a new scale after receiving a manuscript. Instead of relying on quantitative assessment, reviewers can therefore only judge the redundancy of a new scale based on their knowledge of existing scales.

Next to correlational analyses, expert judgements of the incremental value of a new scale is an essential part of the review process. Qualitative evaluations of scale content and overlap

bring advantages, as they neither require participant data nor numerical cut-off values. However, some problems with pure expert assessment remain. Scales might still escape the awareness of experts; for example, if they were published for constructs with different labels or in different research fields (Hagger, 2014; Judge, Erez, Bono, & Thoresen, 2002). Further, automated language processing methods frequently outperform humans in detecting construct similarities (Larsen & Bong, 2016). Thus, automatic semantic analyses are suitable to support expert evaluations in finding and evaluating scale similarities. That being said, we emphasize that computations of semantic overlap replace neither correlational nor expert assessments of scale redundancy. Rather, semi-automated semantic analyses combine the efficiency of quantitative methods and the theoretical grounding of expert evaluations to address limitations of complementary methods. Correlational analyses of convergent and discriminant validity, and expert judgement remain a necessary part of investigating redundancy and incremental value of new scales.

In summation, current methods to protect existing scales from redundant publications come with a number of limitations, which may partly explain the continuous proliferation of psychological scales and constructs. The limitations we identified here are, in short: 1) Researchers are limited in their awareness, resources, and potentially their motivation to collect all data required to properly judge the incremental value of a new scale, 2) diagnosing incremental value or redundancy based on correlation scores alone may endanger theoretical justification of scale content; and 3) reviewers typically cannot collect data and have to rely on their subjective knowledge of "what is out there" and the analyses presented by the authors.

To address these problems in assessing scale redundancy, we introduce the Semantic Scale Network. The Semantic Scale Network is an easy-to-use online application that supports

traditional methods of scale assessment by detecting and quantifying semantic similarities between a new scale and all the scales in the application's corpus.

Importantly, quantifications of semantic similarities (i.e., similarities in question content between scales) are generated through the application without access to any participant data. The application therefore allows authors and readers alike to spot and evaluate semantic overlap between scales, which will hopefully stimulate discussion about the uniqueness and necessity of new scales. The Semantic Scale Network also addresses the problem that authors and reviewers need to be aware of every potentially related scale, since comprehensive scale repositories, integrated in the application, can automatically detect the scales with the highest semantic similarity. Lastly, conducting semantic analyses of scale redundancy directs attention to item content, which is informative for evaluating questionnaire redundancies beyond correlation scores (Arnulf et al., 2014; Arnulf et al., 2018; Clark & Watson, 1995; Gefen & Larsen, 2017; MacKenzie et al., 2011; Maul, 2017). Reporting insights from this similarity analysis next to established scale criteria (e.g., correlational reliability and validity scores) could become standard procedure for publishing and reviewing a new scale. In the following paragraphs we explain the concept and assessment of semantic similarity including exemplary results of this online application.

Semantic Similarity

At the roots of the application is the computation of semantic similarity. Lin (1998) defined the semantic similarity between two objects as: "the ratio between the amount of information in the commonality and the amount of information in the description of the two objects" (p. 298). Tests of semantic similarity therefore answer the question of how strongly two

texts contain similar (as opposed to dissimilar) content. Translated to psychological scale development the question becomes: To what degree do two scales target the same construct, as opposed to different constructs, based on their question texts?

With this definition of semantic similarity in mind, we can turn our attention to computing semantic similarity. Quantitative analyses of similarity between two documents (e.g., an archived scale/document and a new scale/search query), constitute a substantial part of the research field of information retrieval (for an introduction see Zhai, 2008). Thus, the presented application shares some methodological features (including semantic processing) with well-known search engines on the internet (cf. description of Google Scholar in Beel, Gipp, & Wilde, 2009). Below, we briefly describe basic word matching, before introducing Latent Semantic Analysis (LSA). LSA builds on and improves word matching and is the underlying algorithm of the Semantic Scale Network.

Word matching. Assume we have the following three texts: "I usually enjoy parties", "My manager is cruel", and "I enjoy dress-up parties". Intuitively we recognize that the first and the last text are similar, as both mention enjoyment of parties. Word matching works just like that, by using words appearing in different texts as an indicator of text similarity. The basis for calculating this similarity is a document-term matrix as illustrated in Table 1 that indicates which documents (texts) contain which terms (words).

Insert Table 1 about here

We can see that text 1 and text 3 are relatively similar because three words match between both texts (highlighted in gray in Table 1) whereas neither of these texts share words

with text 2. These matching scores can be translated into a cosine similarity that, if we interpret word scores as coordinates, indicates the closeness of the texts in a space with as many dimensions as we have words in the texts. This cosine similarity can take on values from 0 to 1, with 1 indicating identical texts (all words/coordinates match) and 0 indicating no textual overlap. For example, to compare text 1 and 3, the cosine similarity is given by

$CS(v_1, v_2) = \frac{v_1^T v_3}{\| v_1 \| \| v_3 \|} = 0.75$, where v_1 is the row-vector from Table 1 that corresponds to text 1, and v_3 is the equivalent vector for text 3. Since all entries in these vectors are 1 or 0, their product will be a sum of shared words, i.e. $v_1^T v_3 = 3$. The formula further indicates that this vector multiplication is divided by the corresponding vector lengths, with $\| v_1 \|$ being equal to $\sqrt{v_1^T v_1}$. This cosine normalization ensures that all vectors have unit length and the similarities are based on the vector direction in the semantic space rather than the vector length. Put differently, psychological scales with many items and words are more likely to have matching words with other scales regardless of content. The cosine normalization accounts for such inflated similarity scores.

Word importance. When we reason why text 1 and text 3 are related, most of us would highlight that the texts match in their mentioning of "enjoy" and "parties". However, they also overlap in regard to the word "I". Correctly, one might forward that the word "I" is less important, because it is likely to appear in texts about any topic. Thus, it is hardly ever characteristic of a scale, nor does it give insight into scale similarities. In order to account for these differences in term importance, the word scores are weighted by the frequency of their

occurrence in the total corpus. More precisely, researchers usually take the ratio of all documents (i.e., scales) to the number of documents that the word appears in, then compute the logarithm of this ratio, and finally multiply the result with the original word scores (Ramos, 2003). Through this procedure, frequencies of words that appear in almost every text (e.g., "the" in generic texts, or "think" in psychological scales) are shrunk towards zero as they do not serve as strongly to characterize or distinguish texts. This normalization procedure is called term-frequency - inverse document-frequency (tf-idf) normalization. There are alternative forms of these normalization steps in information science, but they are all implemented to serve the same purpose of taking into account differences in word importance before computing a similarity score (Gefen, Endicott, Fresneda, Miller, & Larsen, 2017).

Preprocessing. Aside from word count normalization, preprocessing texts has also proven beneficial for calculating semantic overlap. Basic preprocessing steps include deleting stop-words (e.g., "the", "and", "to"), removing punctuation, and removing numbers. An example for the usefulness of deleting such features is given when looking at psychological scales. Some authors publish their scale items with numbers or letters preceding each item, while some do not. Some authors use colons or periods at the end of each item while others do not. Removing these features helps to focus the similarity computation on what really counts.

Two last powerful preprocessing steps that we briefly describe are lemmatization and stemming, which successively trim down each word to its word stem before computing similarities. Imagine that text 1 had not included the word "parties" but instead the word "partying". Now, text 1 and text 3 would have one match less, even though "parties" and "partying" describe the same concept. Lemmatization relies on existing word dictionaries to convert words to their base form (i.e., both "parties" and "partying" become "party") and

subsequent analyses can correctly identify the overlap between both texts (e.g., Kanis, & Skorkovská, 2010) Stemming is an additional, simple method that cuts down the outputted words from the lemmatization to their word stem. To clarify, lemmatizing "managers" and "managing" returns, respectively, "manager" and "manage", which both result in "manag" after subsequent stemming. Especially for short texts like psychological scales this reduction facilitates overlap detection (Hull, 1996). For the same reason it is also common to convert all words to lowercase.

Although the combination of multiple preprocessing steps from word count normalization to cosine computations can already lead to useful quantifications of text similarity, there is also room for improvement. Imagine the third text had not been "I enjoy dress-up parties", but instead "Social gatherings are fun". Most people would still agree that this text is very similar to text 1 ("I usually enjoy parties"), but now there is not a single word match indicating similarity. Especially for comparing psychological scales, which frequently use different but synonymous items, detecting such latent similarities is key. The next technique, which is implemented in the Semantic Scale Network, was designed to detect such latent overlap between texts.

Latent semantic analysis. In order to move from simple word matching to detecting similarity in meaning, the same preprocessing steps should be completed as described in the previous section (i.e., removing stop-words, lemmatization, stemming, lowercasing, tf-idf normalization). Thus, after carrying out these steps, we again have a document-term matrix similar to that in Table 1, but now with normalized entries instead of raw counts, and slightly different terms due to removal of stop-words ("I" is no longer part of the document-term matrix) and lemmatization/stemming (e.g., "parties" becomes "parti").

The crucial improvement beyond word matching is now to recognize that texts using different words can still talk about similar topics (e.g., "gatherings" and "parties"). In order to

determine whether texts using different words are similar, we need to understand whether their words can be summarized under a shared latent topic (e.g., "social event" in the case of "gathering" and "party"). To this purpose, we condense the document-*term* matrix into a smaller document-*topic* matrix through a method that is closely related to principal component analysis. Generally, the words "gathering" and "party" can be expected to co-occur in texts with a heightened probability as both relate to the latent topic "social event". Further, even in the absence of such *direct co-occurrence*, both "gathering" and "party" can be identified as similar because of their *parallel co-occurrence* with other terms (e.g., "friends", "dance", "talk"). Thus, there is a degree of correlation between both words in texts, which is reflected in the document-term matrix.

Such correlational patterns are frequently investigated in psychological research through principal component analysis (PCA). The assumption is usually that there is a latent phenomenon that explains these correlations. In LSA, the reasoning is that the co-occurrences of words are explained by latent topics; that is, certain words often co-occur (directly or in parallel) because they belong to the same topic (Wolfe & Goldman, 2003). Similar to PCA, we can generate a score for each document (in PCA: participant) on each latent topic (in PCA: latent construct). These new topic scores can then replace the higher-dimensional word scores (in PCA often: item scores), transforming our document-term matrix into a lower-dimensional document-topic matrix. The document-topic matrix thus provides us with a more insightful semantic space where texts can have similar scores on a topic despite using different words (Kjell, Kjell, Garcia, & Sikström, 2018).

The mathematical method for dimensionality reduction in LSA is singular value decomposition (SVD), which is also at the core of PCA. As Figure 1 illustrates, the goal of SVD

in LSA is to express a document-term matrix A as a product of three matrices containing the matrix's eigenvectors and eigenvalues, $A_{[docs,\,terms]} = U_{[docs,\,topics]} * \Sigma_{[topics,\,topics]} * V^{T}{}_{[terms,\,topics]}$.

Insert Figure 1 about here

The rows of matrix U contain the left-singular vectors of the original document-term matrix (i.e., eigenvectors of AA^T) and show how much each document loads on each of the latent topics. Conversely, the rows of matrix V contain the right-singular vectors of the original document-term matrix (i.e., eigenvectors of A^TA) and describe how much each latent topic loads on each of the terms. As such, the matrix V can be used to interpret the content of each latent topic. Finally, the matrix Σ contains the eigenvalues of the original document-term matrix A.

As in PCA, the first k eigenvectors and eigenvalues (from the left in the three matrices U, Σ, and V) capture the highest amount of variance of the original data, and we can therefore truncate the solution and still approximate the original data closely. In the Semantic Scale Network, the truncated matrix U_k therefore describes how strongly each psychological scale loads on each of the k latent topics that explain most semantic variance in the corpus of psychological scales. The truncated matrix U_k can now be utilized to generate cosine similarities as indications for similarities between scales in the same way as was the case for word matching. However, as LSA topic scores can be negative, the cosine similarities now lie between -1 and 1 with negative values indicating dissimilarity (i.e., high distance between texts). A widely cited introduction to LSA is provided by Deerwester and colleagues (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). A tutorial paper for implementing LSA in R is provided by Gefen and colleagues (Gefen, Endicott, Fresneda, Miller, & Larsen, 2017). A review of the

literature shows that various alternatives to LSA exist in the fields of linguistics and computational social science (for a recent review see: Pradhan, Gyanchandani, & Wadhvani, 2015). Still, when put to the test in practical scenarios, many procedures for computing semantic similarity actually give similar results (Mihalcea, Corley, & Strapparava, 2006). We chose LSA over, for instance, latent Dirichlet allocation (Blei, Ng, & Jordan, 2003), because LSA was shown to outperform LDA in the context of psychological single-construct texts (Larsen & Bong, 2016), and because LSA is closer to well-known statistical methods in psychology. Training more advanced models, most prominently neural network architectures, requires much more labeled text data than the current corpus provides (e.g., Altszyler, Sigman, Ribeiro, & Slezak, 2016). However, there are ways to enhance the size of the training set by relying on previous data publications. Thus, we demonstrate in the supplementary materials how large amounts of (non-psychological) text data (Google News corpus) could be used in combination with a pre-trained word2vec model (a shallow neural network; Mikolov, Chen, Corrado, & Dean, 2013) to compute similarities between psychological scales. While a qualitative assessment suggests similar performance of both approaches, LSA and word2vec, future research could generate a large labeled training set to further fine tune and optimize the selection of algorithms. We describe the LSA approach as it is competitive in performance, trained on psychological data (scale texts), and closely related to statistical models in psychology.

The Semantic Scale Network

LSA is the main method underlying the Semantic Scale Network. The main output of the application are quantifications and visualizations of semantic similarity (i.e., cosine similarity) between psychological scales. The Semantic Scale Network was developed as a decision support systems (DSS) to help researchers and reviewers detect related scales, protect existing scales

from redundant publications, and quantify semantic construct overlap (i.e., semantic validity; Larsen, Nevo, & Rich, 2008). As a DSS, the Semantic Scale Network falls into "the area of the information systems (IS) discipline that is focused on supporting and improving […] decision making" (Arnott & Pervan, 2014). In recent years, psychological science has adopted multiple DSS's, often to improve research and publication quality (e.g., "statcheck" by Epskamp & Nuijten, 2016). Through the current work, we aim to contribute a DSS that helps improve psychological science by uncovering and preventing scale redundancies. We go on to describe the application's scale corpus, example outputs of the application, and best practices regarding the interpretation of results.

Data

Many psychological scales that we included in the application's corpus were found in public questionnaire repositories. We included openly-accessible scales from the International Personality Item Pool (ipip.ori.org; Goldberg et al., 2006), the Measurement Instrument Database for the Social Sciences (midss.org; Whitaker Institute for Innovation and Social Change , n.d.), the Registery for Scales and Measures (scalesandmeasures.net; Santor, 2013), Ron Okada's collection of scales for students (yorku.ca/rokada/psyctest; Okada, 2018), the Association of Religion Data Archives (thearda.com/mawizard/scales; The Association of Religion Data Archives, n.d.), the Open Source Psychometrics Project (openpsychometrics.org/_rawdata; Open Psychometrics, n.d.), the Longitudinal Internet Studies for the Social sciences (dataarchive.lissdata.nl/concepts; CentErdata, n.d.), the Inter-Nomological Network (https://inn.theorizeit.org/; Human Behavior Project, 2011), psychology tools (psychologytools.com; Psychology Tools, 2018), the Positive Psychology Center of the University of Pennsylvania (ppc.sas.upenn.edu/resources/questionnaires-researchers; Schulman,

n.d.), and the scale collection of the decision lab (decisionlab.shinyapps.io/InterindividualDifferenceMeasures; Fiedler & Lyubenova, in preparation). We discarded restricted access publications, because users of the application need access to the scale items in order to interpret the application's output. Further, we added many scales that were submitted by researchers in psychology following open calls over social media, mailing lists, and personal communication. Scale submissions can be made at any time on the application's website. Any scale with item texts, which is published in a peer-reviewed journal and of which the items are freely available online, qualifies for inclusion. The corpus of scales in continuously growing, and contained 4,037 scales at the time of submitting this paper. The current number can be found on the application's website. All included scales can be accessed through references presented on the application's website.

The distribution of cosine similarities between all included scales ranges from -.66 to 1. The corpus includes pairs of scales with cosine = 1 (i.e., perfectly redundant scales), simply because some scales for assessing a construct for different populations are actually comprised of the same items. The average cosine similarity between two scales is .007 (*Mdn* = -.001, *SD* = .076). It is not surprising that this constitutes a very small similarity given the wide spectrum of constructs that are assessed by the corpus. More interesting in terms of redundancy and overlap is therefore the distribution of cosine similarities between a scale and the scale to which it is most closely related (i.e., its closest neighbor in the network). Here, we find that the average cosine similarity is .68 (*Mdn* = .674, *SD* = .164). The distributions are depicted in Figure 2.

Insert Figure 2 about here

When using the application, we advise the reader to always evaluate the similarities of the entered scale with its closest neighbors regardless of the returned cosine similarity value. Although a numerical rule of thumb when deciding whether two scales overlap 'very much' or 'only marginally' might seem appealing, we believe that generic cut-offs are not well suited for the current application, because they are often misleading (Lakens et. al, 2018) and could inhibit a proper examination of item content. Further, as the semantic corpus of psychological scales is continuously expanding, cosine values might shift slightly and thereby cross arbitrary cut-off lines. To guide users, we therefore provide an example of how to investigate the semantic similarity of a new scale with scales in the application's corpus.

Example Results

The most basic and for most of us most useful feature of the Semantic Scale Network is to identify overlap between two scales in order to discuss a scale's incremental value. Imagine encountering a new psychological scale. We assume that the scale passes established criteria for psychological scales (e.g., answer reliability and validity). The hypothetical scale was developed to assess a person's 'social drive' and its items are:

1. I avoid social interaction as much as possible.
2. I think parties are fun.
3. I am outgoing.
4. I have a lot of friends.
5. I rarely enjoy group activities.
6. I like being alone.

After entering the items into the application, the returned results describe the semantic embeddedness of the scale in the scale corpus as depicted in Figure 3.

Insert Figure 3 about here

As indicated in the result table of Figure 3, the new scale appears most strongly related to scales about extraversion and sociability. The cosine similarities lie between .387 and .58. Yet, more important than these numerical values are the item contents of the detected scales. Examining the item contents (under the provided links) allows us to answer the question of *why* the scales are suggested to be semantically similar. For example, some items of a sociability scale (Goldberg et al., 2006), our scale's closest neighbor in the network, are:

1. Usually like to spend my free time with people.
2. Talk to a lot of different people at parties.
3. Love to chat.
4. Make friends easily.
5. Enjoy being part of a group.
6. Rarely enjoy being with people.

In this case, almost all items in our new 'social drive' scale are very closely related with an item in the sociability scale. The marginal uniqueness of our new scale seems to lie primarily in item 6 which addresses enjoyment of being alone. This concept is not directly included in the sociability scale. The insights gained from inspecting the two scales raise three important questions. First, is the incremental value of the new scale sufficient to justify the publication and use of the new scale? Second, is it the concept of 'enjoyment of being alone' what distinguishes the concepts of extraversion and social drive? Third, is this concept addressed by one of the

many other related scales? A next step to determine usefulness vs redundancy of our scale would be to continue the investigation of semantic overlap with its second closest neighbor, the 'enjoyment (expected)' scale. These steps to systematically investigate semantic overlap of a new scale with existing scales integrate expert judgements of redundancy with the application's functionality. Further, it is possible to conduct such semantic analyses before any data is collected, so correlational analyses could "follow up" on analyses of semantic overlap.

Limitations and How Not to Use the Semantic Scale Network

As the application is based on methods that are not commonly used in psychological research it is crucial to discuss the limitations of the application and how not to interpret its output.

Corpus completeness. It is important to realize that the application's corpus, albeit being of substantial and increasing size, will likely never include all scales ever developed. This has two important consequences: First, redundancies of a new scale with already existing scales are not highlighted by the application, if the relevant scales are not in the corpus. Second, the computation of the semantic space in which the psychological scales lie is based on an incomplete language sample. This means that the LSA procedure likely does not capture all semantic topics that can be found in psychological measurement tools. It is for example likely that smaller research fields (e.g., back pain) are only captured with very few or no latent topics. In order for the Semantic Scale Network to evolve into the most useful tool it can be, a collective effort is needed to enlarge the scale corpus. To facilitate this, published scales can be submitted to the corpus under a link provided on the application's website.

Confirmation of uniqueness. Further, it is crucial to realize that the application's functionality can be used to detect potential redundancies, but never to confirm uniqueness. Entering a scale and finding no considerable overlap with any neighbor scale cannot be seen as proof of uniqueness or necessity to add the scale to the existing body of psychological scales. One reason is the aforementioned incompleteness of the application's corpus, which might prevent redundancy detection. Another reason is that the utilization of latent topics improves similarity detection, but it does not perfect it. This means that it is still possible to generate a scale which is closely related to existing scales without using the same words, and without words clustering together in any latent topic. In fact, we believe it is possible for almost any scale to adjust its wording until no strong semantic similarities to other scales can be found. Some strategies and best-practices can alleviate this concern.

First, scale semantics with artificially low similarities will be characterized by unnecessarily rare words. For instance "have a lot of friends" can be rewritten as "have plentiful companions". Reviewers should question the use of words that are obvious synonyms to more intuitive words, not just to ensure similarity detection, but also to facilitate a good understanding of scale items among participants. The Semantic Scale Network may assist reviewers here, as it is possible to actively look for hidden neighbors in the network by replacing individual words with their synonyms before entering the items into the application. Second, the questionable practice to mask similarities with existing scales through the adjustment of words may often deteriorate traditional evaluation criteria for scale development, such as factor structure, Cronbach's alpha, or convergent validity with other scales. Uniqueness hacking therefore becomes impractical. Third, and most importantly, the Semantic Scale Network supports but does not replace expert knowledge about existing scales. Reviewing and citing existing literature

should, for instance, always form part of the examination of item content. For that reason, it is clearly necessary to familiarize oneself with the relevant literature, for instance through general research search engines, or measurement-specific repositories like INN (Human Behavior Project, 2011), or the subscription-based psycTESTS (e.g., Swogger, 2013), and never solely rely on the Semantic Scale Network, when judging uniqueness and incremental value. Such other online tools have, for instance, the advantage of including non-textual (e.g., image-based) scales and tests.

Reliance on numeric similarity. A final caveat is that, as mentioned above, cosine similarities can give a false sense of the 'exactness' of content overlap. For this reason it is not advisable to judge uniqueness and redundancy of a new scale based on a cut-off or rule of thumb. Cosine values usually shift after changing one item in the scale or just a single word, especially if the scale is relatively short. Therefore, the application should be used as a guide to find and investigate similar scales based on their item content. An expert's discussion of item content across scales always provides stronger arguments than high or low cosine values. We therefore advise to always investigate a scale's closest neighbors regardless of their specific cosine value.

Using the Scale Corpus for Original Research

Whereas the focus of the Semantic Scale Network lies on highlighting scale overlap, an important output of the current project is the semantic space of psychological scales. As a whole the application can be understood as a semantic network of questionnaire texts where connection strength is given by semantic overlap. By using the questionnaire texts as empirical data, typical network analyses can be conducted, such as identifying highly centralized scales and scales bridging different scale clusters. Generally, such analyses tap into the hierarchical nature and connectedness of psychological constructs (cf., Judge et al., 2002). We hope that future meta-

analytical research will make use of semantic network analyses and the provided data, for instance, to discuss possibilities of condensing scale clusters into overarching constructs (cf. Hodson et al., 2018). Figure 4 depicts a snippet of the semantic scale network.

Insert Figure 4 about here

Yet another usage of the application's corpus is to examine the latent topics of psychological measurement that can be generated by condensing the document-term matrix. Figure 5 depicts some of these topics as word clouds, alongside the psychological scales that relate most strongly to these topics. Analyzing such latent communalities of psychological measures can give concise summaries of construct clusters and support item development.

Insert Figure 5 about here

Conclusion

Redundant scales lead to arbitrariness and disorientation in psychological measurement, weak theories, and confusion among researchers and practitioners. In order to help controlling scale and construct proliferation we created the Semantic Scale Network—a corpus-based, easy-to-use online application that enables users to find semantically similar psychological scales. The application assists researchers and reviewers in detecting existing scales before redundant scales are published and used.

When using the application for scale comparisons, we discourage relying on generalized cut-off scores in assessing potential redundancy, and highlight the need for expert evaluations of the semantically most similar scales. Such an explorative approach is affordable as researchers do not have to collect data for the examined scales. Further, the application can be used by

researchers to search for relevant scales not based on construct names, but item content. This will allow researchers to find relevant scales even if they were published under unintuitive names and in different research fields.

Aside from research focusing directly on scale development, researchers are free to use the Semantic Scale Network as language input for their own research. The scale corpus serves as an increasingly comprehensive semantic space that captures the language of psychological measurement and can be used for a wide range of language-based research projects (Chen & Wojcik, 2016). As semantic overlap and answer correlations often approximate each other, the network can, for example, be used to investigate which observed correlations between constructs might be grounded exclusively in similar question phrasing. To illustrate, Arnulf and colleagues succeeded to predict between 54% and 86% of survey covariance based on semantic similarity alone (2014). A very different follow-up would be to generate networks for scales in other languages to test whether inconsistent construct correlations observed in different cultures could be explained by inconsistent semantic embeddings of scales (network structure differs between languages).

Open Call to Submit Scales

The Semantic Scale Network is highly dependent on a comprehensive corpus of psychological scales. Although we accumulated sufficient scales from open repositories to capture a large part of the psychological landscape, we realize that there are many scales yet to be included in the corpus. In order to turn the Semantic Scale Network into the best tool that it can be for psychological science, a collective effort is needed. Therefore, we hope that authors and users of peer-reviewed and openly-accessible scales will continue to submit scales to the

Semantic Scale Network and encourage their colleagues to do the same. There are many reasons to submit a scale, among them:

- protect existing scales from redundant scales in the future
- increase visibility and reuse of existing scales
- contribute to an open, free sharing of tools
- improve the application's performance by ensuring good scale coverage
- improve the application's performance by enlarging the sample used for LSA
- contribute to a parsimonious body and collective maintenance of scale measures

We hope that, the Semantic Scale Network can help prevent further development of redundant psychological scales. Ultimately, this should help psychologists test and discuss theories more correctly and efficiently.

Acknowledgements

The Semantic Scale Network was conceived at the Summer Institute in Computational Social Science (SICSS) in Helsinki, 2018. We thank Matti Nelimarkka, Juho Pääkkönen, Pihla Toivanen, Anders Grundtvig, Robin Lybeck, and other participants of SICSS for their support and comments at the start of this project. Further, we thank Anthony Evans, Marcel Zeelenberg, Joost van Baal-Ilić, Michèle Nuijten, Kai Larsen, Jan-Ketil Arnulf, and Isabel Thielman for their help and advice. We also thank all those who have contributed scales to the Semantic Scale Network so far, as well as those who will contribute in the future.

References

Altszyler, E., Sigman, M., Ribeiro, S., & Slezak, D. F. (2016). Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. *arXiv preprint arXiv:1610.01520.*

Arnott, D., & Pervan, G. (2014). A critical analysis of decision support systems research revisited: the rise of design science. *Journal of Information Technology, 29*(4), 269-293.

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

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

Attali, D. (2018). shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds. R package version 1.0. https://CRAN.R-project.org/package=shinyjs

Bail, C. A. (2016). Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *Proceedings of the National Academy of Sciences, 113*, 11823-11828.

Bailey, E. (2015). shinyBS: Twitter Bootstrap Components for Shiny. R package version 0.61. https://CRAN.R-project.org/package=shinyBS

Banks, G. C., McCauley, K. D., Gardner, W. L., & Guler, C. E. (2016). A meta-analytic review of authentic and transformational leadership: A test for redundancy. *The Leadership*

Quarterly, *27*, 634-652.

Bar-Anan, Y., & Vianello, M. (2018). A multi-method multi-trait test of the dual-attitude perspective. *Journal of Experimental Psychology: General, 147*(8), 1264-1272.

Beel, J., Gipp, B., & Wilde, E. (2009). Academic Search Engine Optimization (aseo): Optimizing scholarly literature for Google Scholar & Co. *Journal of Scholarly Publishing, 41*(2), 176-190.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3, 774-778.

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

Boutaris, T., & Zauchner, C. (2017). tableHTML: A tool to create HTML tables. R package version 1.1.0. https://CRAN.R-project.org/package=tableHTML

Bruner, G. C. (2003). Combating scale proliferation. *Journal of Targeting, Measurement and Analysis for Marketing*, *11*(4), 362-372.

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

CentErdata (n.d.). *Longitudinal Internet Studies for the Social Sciences*. Retrieved from https://dataarchive.lissdata.nl/concepts

Chang, W. (2018). shinythemes: Themes for Shiny. R package version 1.1.2. https://CRAN.R-project.org/package=shinythemes

Chang, W., Cheng, J., Allaire, J.J., Xie, Y., & McPherson, J. (2018). shiny: Web Application Framework for R. R package version 1.1.0. https://CRAN.R-project.org/package=shiny

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

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*, 309-319.

Cole, M. S., Walter, F., Bedeian, A. G., & O'Boyle, E. H. (2012). Job burnout and employee engagement: A meta-analytic examination of construct proliferation. *Journal of Management*, *38*, 1550-1581.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391-407.

Dowle, M., & Srinivasan, A. (2018). data.table: Extension of `data.frame`. R package version 1.11.8. https://CRAN.R-project.org/package=data.table

Epskamp, S., Cramer, A. O., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software, 48*, 1-18.

Epskamp, S., & Nuijten, M. B. (2016). *statcheck: Extract statistics from articles and recompute p values*. R package version 1.2.2. http://CRAN.R-project.org/package=statcheck.

Fiedler, S. & Lyubenova, A. (in preparation). *Individual difference measures in the context of economic decision making: Introducing the DecisionLab Indvidual Differences*

Database.

Gefen, D., Endicott, J., Fresneda, J., Miller, J., & Larsen, K. R. (2017). A guide to text analysis with latent semantic analysis in R with annotated code studying online reviews and the Stack Exchange community. *Communications of the Association for Information Systems*, *41*, 450-496.

Gefen, D., & Larsen, K. (2017). Controlling for lexical closeness in survey research: A demonstration on the technology acceptance model. *Journal of the Association for Information Systems*, *18*, 727-757.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40, 84-96.

Hagger, M. S. (2014). Avoiding the "déjà-variable" phenomenon: Social psychology needs more guides to constructs. *Frontiers in Psychology*, *5*, 52.

Haynes, S. N., & Lench, H. C. (2003). Incremental validity of new clinical assessment measures. *Psychological Assessment*, *15*, 456-466.

Hodson, G., Book, A., Visser, B. A., Volk, A. A., Ashton, M. C., & Lee, K. (2018). Is the Dark Triad common factor distinct from low Honesty-Humility?. *Journal of Research in Personality, 73*, 123-129.

Hull, D. A. (1996). Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science, 47*(1), 70-84.

Human Behavior Project. (2011). Inter-Nomological network. Retrieved April 15, 2019, from

https://inn.theorizeit.org/

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: conceptual, methodological, and statistical issues. *Psychological Assessment*, *15*, 446-455.

Judge, T. A., Erez, A., Bono, J. E., & Thoresen, C. J. (2002). Are measures of self-esteem, neuroticism, locus of control, and generalized self-efficacy indicators of a common core construct? *Journal of Personality and Social Psychology*, *83*, 693-710.

Kanis, J., & Skorkovská, L. (2010, September). Comparison of different lemmatization approaches through the means of information retrieval performance. In *International Conference on Text, Speech and Dialogue* (93-100). Springer, Berlin, Heidelberg.

Kjell, O. N., Kjell, K., Garcia, D., & Sikström, S. (2018). Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods*. Advance online publication. http://dx.doi.org/10.1037/met0000191

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., ... & Buchanan, E. M. (2018). Justify your alpha. *Nature Human Behaviour*, *2*, 168.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: what did they really say?. *Organizational research methods, 9*(2), 202-220.

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

Larsen, K. R., Nevo, D., & Rich, E. (2008). Exploring the semantic validity of questionnaire scales. In *Proceedings of the Annual Hawaii International Conference on System Sciences*.

Le, H., Schmidt, F. L., Harter, J. K., & Lauver, K. J. (2010). The problem of empirical redundancy of constructs in organizational research: An empirical investigation. *Organizational Behavior and Human Decision Processes*, *112*, 112-125.

Lin, D. 1998. An information-theoretic definition of similarity. *Proceedings of the 15 th ICML*, 296-304, Madison, WI.

Loewenthal, K., Lewis, C. (2001). *An introduction to psychological tests and scales.* London: Psychology Press.

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*, 293-334.

Maechler, M., Davis, T.A., Oehlschlägel, J., & Riedy, J. (2018). Matrix: Sparse and Dense Matrix Classes and Methods. R package version 1.2-14. http://matrix.r-forge.r-project.org/

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives*, *15*, 51-69.

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *Proceedings of the 21st National Conference on Artificial Intelligence*, 775–780.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

Morrow, P. C. (1983). Concept redundancy in organizational research: The case of work

commitment. *Academy of Management Review*, *8*, 486-500.

Muis, K. R., Duffy, M. C., Trevors, G., Ranellucci, J., & Foy, M. (2014). What were they thinking? Using cognitive interviewing to examine the validity of self-reported epistemic beliefs. *International Education Research, 2*(1), 17-32.

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

Okada, R. (2018). *Psychological tests for student use*. Retrieved from http://www.yorku.ca/rokada/psyctest/

Olaru, G., Schroeders, U., Hartung, J., & Wilhelm, O. (in press). Ant Colony Optimization and Local Weighted Structural Equation Modeling. A Tutorial on Novel Item and Person Sampling Procedures for Personality Research. *European Journal of Personality*.

Open Psychometrics (n.d.). *Raw data from online personality tests*. Retrieved from https://openpsychometrics.org/_rawdata/

Pekrun, R., Vogl, E., Muis, K. R., & Sinatra, G. M. (2017). Measuring emotions during epistemic activities: the Epistemically-Related Emotion Scales. *Cognition and Emotion*, *31*, 1268-1276.

Pradhan, N., Gyanchandani, M., & Wadhvani, R. (2015). A review on text similarity technique used in IR and its application. *International Journal of Computer Applications*, *120*, 29-34.

Psychology Tools (2018). *Psychological Scales and Measures.* Retrieved from https://www.psychologytools.com/download-scales-and-measures/

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. In *First International Conference on Machine Learning* (133-142). New Brunswick: Rutgers University.

Reeve, C. L., & Basalik, D. (2014). Is health literacy an example of construct proliferation? A conceptual and empirical evaluation of its redundancy with general cognitive ability. *Intelligence*, *44*, 93-102.

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

Roodt, G. (2004). Concept redundancy and contamination in employee commitment research: Current problems and future directions. *SA Journal of Industrial Psychology*, *30*(1), 82-90.

RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. http://www.rstudio.com/.

Santor, D. A. (2013). *Registry of scales and measurements*. Retrieved from http://scalesandmeasures.net

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological assessment* (E-publication ahead of print).

Schulman, P. (n.d.). *Positive Psychology Center's questionnaires for researchers.* Retrieved from https://ppc.sas.upenn.edu/resources/questionnaires-researchers

Sechrest, L. (1963). Incremental validity: A recommendation. *Educational and Psychological Measurement*, *23*(1), 153-158.

Shaffer, J. A., DeGeest, D., & Li, A. (2016). Tackling the problem of construct proliferation: A guide to assessing the discriminant validity of conceptually related constructs. *Organizational Research Methods, 19*, 80–110.

Swogger, S. E. (2013). PsycTESTS. *Journal of the Medical Library Association: JMLA, 101*(3), 234-235.

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., ... & de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*(1), 34-42.

The Association of Religion Data Archives (n.d.). *Multi-item measures.* Retrieved from http://thearda.com/mawizard/scales/

Whitaker Institute for Innovation and Social Change (n.d.). *Measurement Instrument Database for the Social Sciences.* Retrieved from http://www.midss.org

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis.* New York: Springer.

Wickham, H., Francois, R., Henry, L., & Müller, K. (2018). dplyr: A Grammar of Data Manipulation. R package version 0.7.8. https://CRAN.R-project.org/package=dplyr

Wickham, H., Hester, J., & Francois, R. (2017). readr: Read Rectangular Text Data. R package version 1.1.1. https://CRAN.R-project.org/package=readr

Wild, F. (2015). lsa: Latent Semantic Analysis. R package version 0.73.1. https://CRAN.R-project.org/package=lsa

Wolfe, M. B., & Goldman, S. R. (2003). Use of latent semantic analysis for predicting psychological phenomena: Two issues and proposed solutions. *Behavior Research Methods, Instruments, & Computers*, *35*(1), 22-31.

Xie, Y., Cheng, J., & Tan, X. (2018). DT: A Wrapper of the JavaScript Library 'DataTables'. R package version 0.5. https://CRAN.R-project.org/package=DT

Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of Research in Personality, 44*(2), 180-198.

Zhai, C. (2008). Statistical language models for information retrieval. *Synthesis Lectures on Human Language Technologies, 1*(1), 1-141.

Table 1

Document-term matrix for the three example texts.

	Term								
Document	i	usually	enjoy	parties	dress-up	my	manager	is	cruel
text 1	1	1	1	1	0	0	0	0	0
text 2	0	0	0	0	0	1	1	1	1
text 3	1	0	1	1	1	0	0	0	0

Note. Number of times a term (column) occurs in a document (row). Overlap between text 1 and text 3 is highlighted in gray.

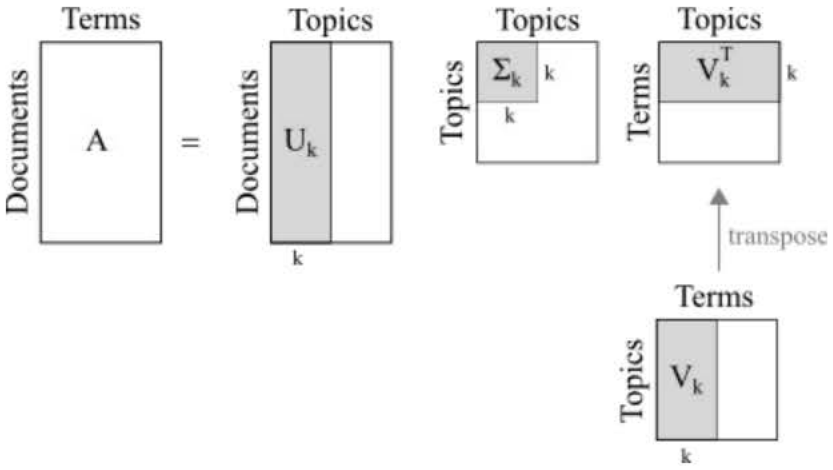

Figure 1. Illustration of a singular value decomposition (SVD) in the context of latent semantic analysis (LSA). Closely based on Fig. 2.1 in Martin & Berry (2007).

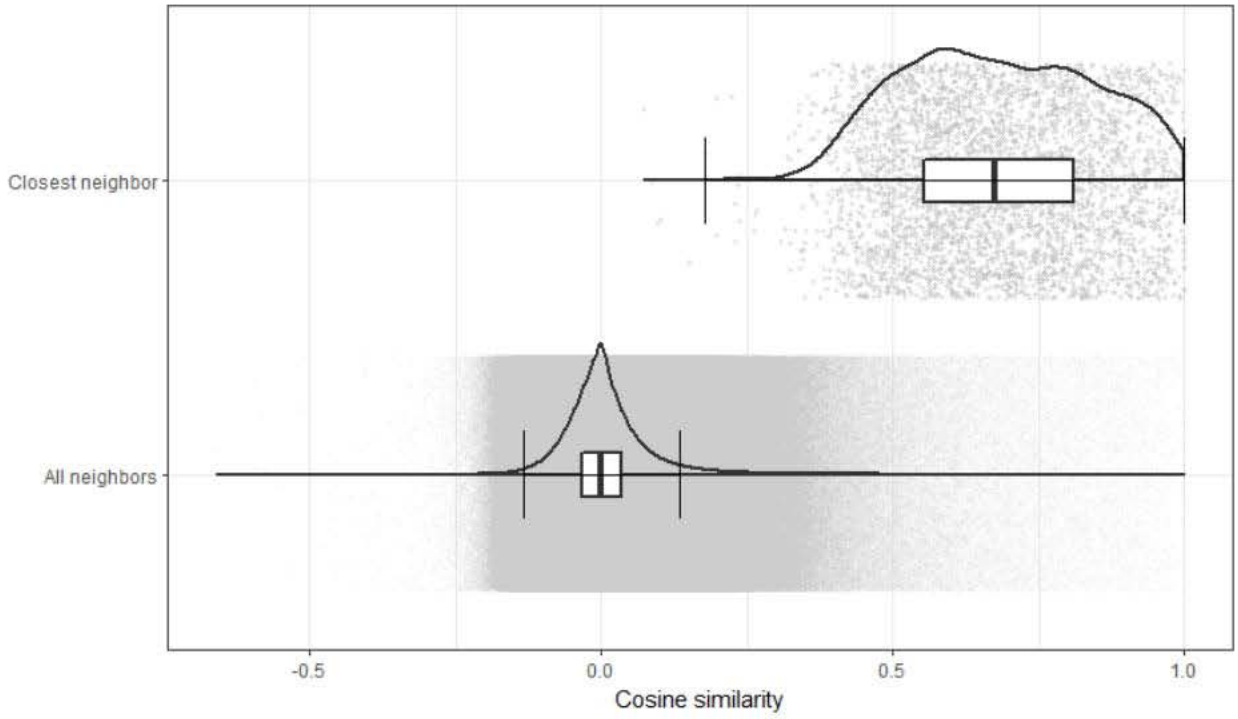

Figure 2. Violin plots of cosine similarities between each scale and its most similar neighbor (top) and similarities between all pairs of scales (bottom).

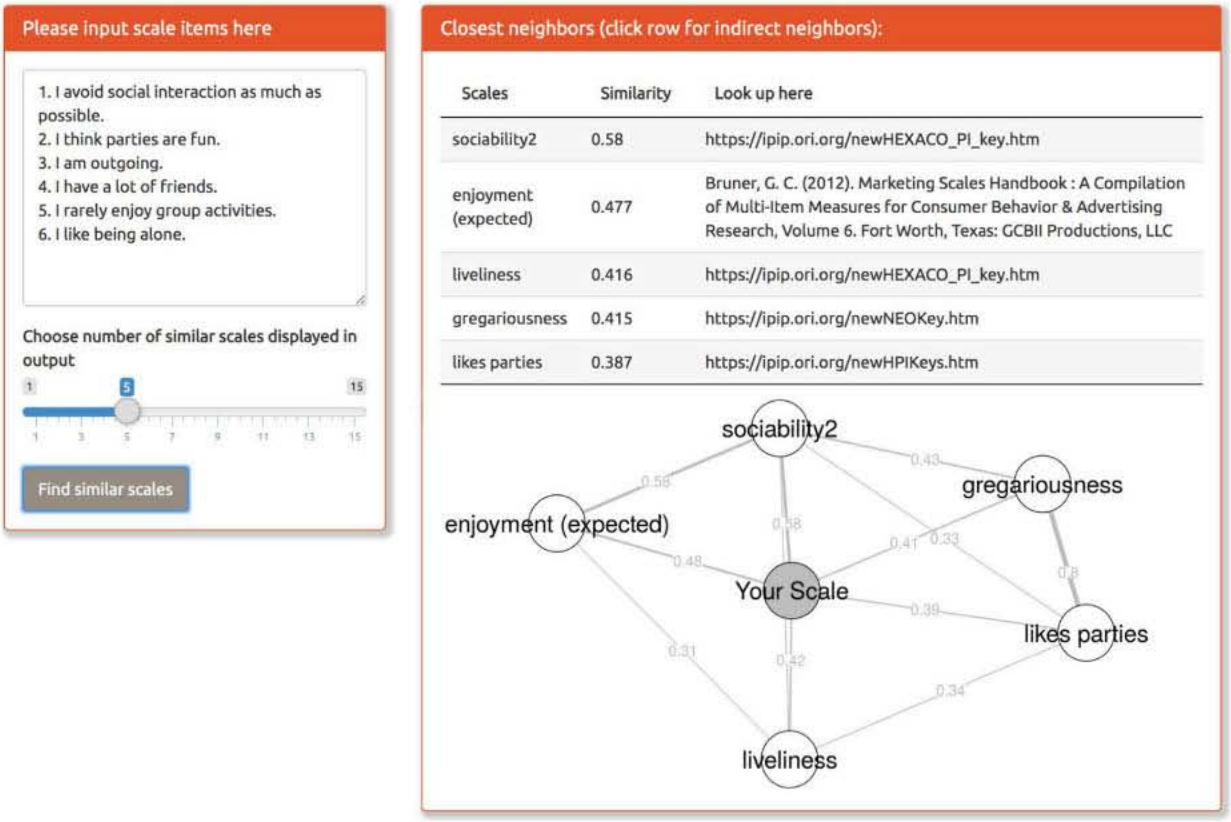

Figure 3. Screenshot of the application's output.

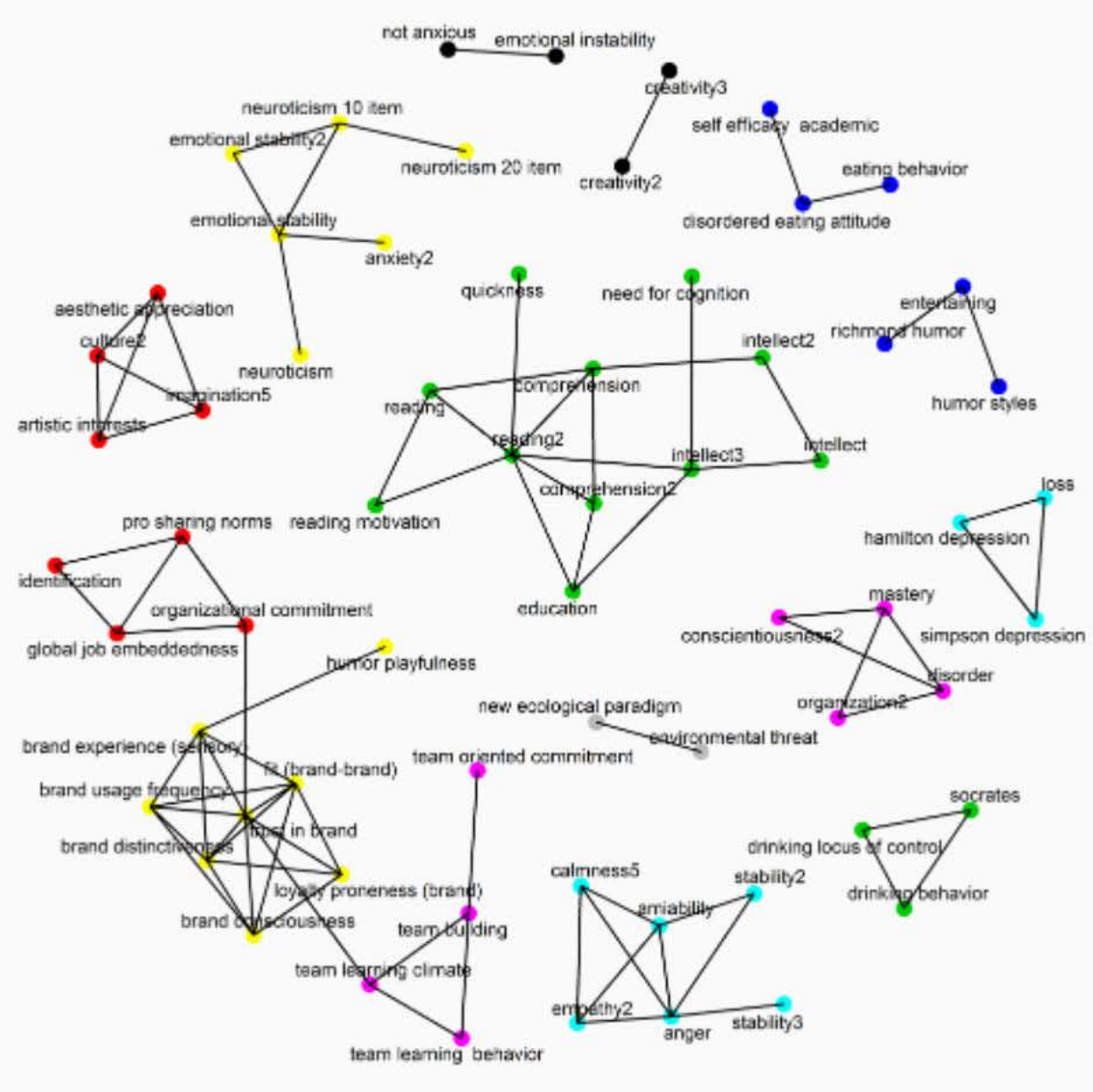

Figure 4. Small snippet of the Semantic Scale Network. Each node is a psychological scale and the edges are drawn based on semantic similarities (i.e. cosine similarities).

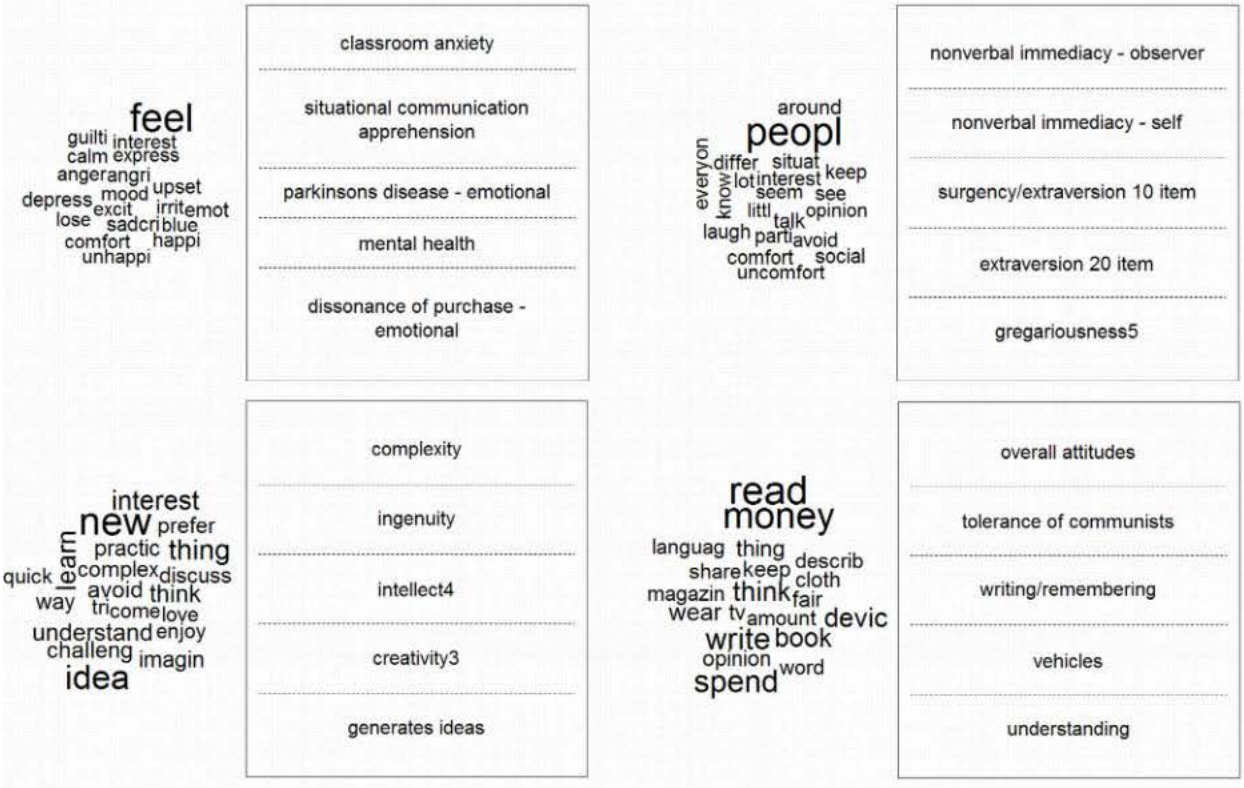

Figure 5. Latent topics are plotted as word clouds. The words in each cloud are characteristic of this topic. On the right side of each topic are the psychological scales that score highest on this latent topic. Many topics have an intuitive interpretation that we can assign to them. For instance, we would relate the top-left topic to emotion, bottom-left to cognition, and top-right to social behaviors. Many other topics are not easily interpretable as they might not pertain to specific psychological constructs. The bottom-right topic, for instance, seems to capture a mix of constructs relating to money, products, and reading.

Appendix

R packages used

All data manipulations and analyses were done in R Studio (RStudio Team, 2016) using the language R (R Core Team, 2018). In order to read in and manipulate the text data we used the packages 'readr' (version 1.1.1; Wickham, Hester, & Francois, 2017), 'data.table' (version 1.11.8; Dowle, & Srinivasan, 2018), 'dplyr' (version 0.7.8; Wickham, Francois, Henry, & Müller, 2018), and 'textstem' (version 0.1.4; Rinker, 2018). The packages 'lsa' (version 0.73.1; Wild, 2015), 'quanteda' (version 1.4.3; Benoit et al., 2018), 'tableHTML' (version 1.1.0; Boutaris, & Zauchner, 2017), 'qgraph' (version 1.5; Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012), ggplot2 (version 2.2.1; Wickham, 2009), and 'textnets' (version 0.1.1; Bail, 2016) were used to analyze and visualize the data. The online application 'The Semantic Scale Network' was created and designed using the packages 'shiny' (version 1.2.0; Chang, Cheng, Allaire, Xie, & McPherson, 2018), 'shinyjs' (version 1.0; Attali, 2018), 'shinyBS' (version 0.61; Bailey, 2015), and 'shinythemes' (version 1.1.2; Chang, 2018), and optimized using the packages 'DT' (version 0.5; Xie, Cheng, & Tan, 2018), and 'Matrix' (version 1.2-14; Maechler, Davis, Oehlschlägel, & Riedy, 2018).