American Psychological Association

Psychological Methods

1082-989X/19/$12.00

2019, Vol. 24, No. 1, 92–115
http://dx.doi.org/10.1037/met0000191

Semantic Measures: Using Natural Language Processing to Measure, Differentiate, and Describe Psychological Constructs

Oscar N. E. Kjell and Katarina Kjell
Lund University

Danilo Garcia
Blekinge County Council, Karlskrona, Sweden, and University of Gothenburg

Sverker Sikström
Lund University

Abstract

Psychological constructs, such as emotions, thoughts, and attitudes are often measured by asking individuals to reply to questions using closed-ended numerical rating scales. However, when asking people about their state of mind in a natural context ("How are you?"), we receive open-ended answers using words ("Fine and happy!") and not closed-ended answers using numbers ("7") or categories ("A lot"). Nevertheless, to date it has been difficult to objectively quantify responses to open-ended questions. We develop an approach using open-ended questions in which the responses are analyzed using natural language processing (Latent Semantic Analyses). This approach of using open-ended, semantic questions is compared with traditional rating scales in nine studies (N = 92–854), including two different study paradigms. The first paradigm requires participants to describe psychological aspects of external stimuli (facial expressions) and the second paradigm involves asking participants to report their subjective well-being and mental health problems. The results demonstrate that the approach using semantic questions yields good statistical properties with competitive, or higher, validity and reliability compared with corresponding numerical rating scales. As these semantic measures are based on natural language and *measure*, *differentiate*, and *describe* psychological constructs, they have the potential of complementing and extending traditional rating scales.

Translational Abstract

We develop tools called semantic measures to statistically measure, differentiate and describe subjective psychological states. In this new method, natural language processing is used for objectively quantifying words from open-ended questions, rather than the closed-ended numerical rating scales traditionally used today. Importantly, the results suggest that these semantic measures have competitive, or higher, validity and reliability compared with traditional rating scales. Using semantic measures also brings along advantages, including an empirical description/definition of the measured construct and better differentiation between similar constructs. This method encompasses a great potential in terms of improving the way we quantify and understand individuals' states of mind. Semantic measures may end up becoming a widespread alternative applied in scientific research (e.g., psychology and medicine) as well as in various professional contexts (e.g., political polls and job recruitment).

Keywords: psychological assessment, natural language processing, latent semantic analyses

Supplemental materials: http://dx.doi.org/10.1037/met0000191.supp

This article was published Online First July 2, 2018.

Oscar N. E. Kjell and Katarina Kjell, Department of Psychology, Lund University; Danilo Garcia, Blekinge Center of Competence, Blekinge County Council, Karlskrona, Sweden, and Department of Psychology, University of Gothenburg; Sverker Sikström, Department of Psychology, Lund University.

Parts of this article were presented at the 31st International Congress of Psychology in Yokohama, Japan, July 24–29, 2016. This research is supported by a grant from the Swedish Research Council (2015-01229). Oscar N. E. Kjell and Sverker Sikström conceived the studies on reports regarding external stimuli and subjective states of harmony and satisfaction; Katarina Kjell, Oscar N. E. Kjell, and Sverker Sikström conceived the studies on reports regarding subjective states of depression and worry. Oscar N. E. Kjell and Katarina Kjell collected the data. Oscar N. E. Kjell, Katarina Kjell, and Sverker Sikström analyzed the data. Sverker Sikström developed the software for analyzing the parts using natural language processing. Oscar N. E. Kjell, Katarina Kjell, Danilo Garcia, and Sverker Sikström wrote the article. The authors have no competing interest to declare.

Correspondence concerning this article should be addressed to Oscar N. E. Kjell or Sverker Sikström, Department of Psychology, Lund University, Institutionen för psykologi, Box 213, 221 00 LUND. E-mail: Oscar.Kjell@psy.lu.se or Sverker.Sikstrom@psy.lu.se

Rating scales is the dominant method used for measuring people's mental states, and is widely used in behavioral and social sciences, but also in practical applications. These scales consist of items such as *I am satisfied with my life* (Diener, Emmons, Larsen, & Griffin, 1985) coupled with predefined response formats, often ranging from 1 = *strongly disagree* to 7 = *strongly agree* (Likert, 1932). Hence, this method requires one-dimensional, closed-ended answers from respondents. These methods furthermore require the participants to perform the cognitive task of translating their mental states, or natural language responses, into the one-dimensional response format to make them fit current methods in behavioral sciences. In contrast, we argue that future methods in the field should be adapted to the response format used by people, where natural language processing and machine learning may solve the problem of translating language into scales. In summary, the burden of translating mental states into scientific measurable units should be placed on the method, not the respondents. Furthermore, this method conveys limited information concerning respondents' states of mind, as their options for expressing themselves are limited. For example, an individual answering "7" here indicates a high level of satisfaction, but there is no information as to *how* the respondent has interpreted the item or upon which aspects he or she based the answer: Did the individual consider his or her financial situation, relationships with others, both or perhaps something entirely different?

Although numerical rating scales are widespread, easily quantifiable and have resulted in important findings in different fields, they have drawbacks (e.g., see Krosnick, 1999), which are addressed by our approach. Taking advantage of the human ability to communicate using natural language, we propose a method for enabling open-ended responses that are statistically analyzed by means of *semantic measures*. We argue that this method contains several advantages over numerical rating scales. First, in daily life subjective states tend to be communicated with words rather than numbers. A person wanting to find out how their friend feels or thinks tends to allow open-ended answers rather than requiring closed-ended numerically rated responses. However, the rating scale method requires the respondent (rather than the experimental or statistical methods) to perform the mapping of their natural language responses into a one-dimensional scale. Second, the respondent is not forced to answer using (rather arbitrary) numerical rating scales, but is encouraged to provide reflective, open answers. Third, the construct measured is immediately interpreted by respondents, allowing them to freely elaborate on a personally fitting answer. Closed-ended items use a fixed response format imposed by the test developer (e.g., Kjell, 2011), whereas semantic questions (e.g., "Overall in your life, are you satisfied or not?") allow respondents to answer freely concerning what they perceive to be important aspects of a psychological construct. Finally, the semantic measures may also describe the to-be-measured construct; for example, by revealing statistically significant *keywords* describing or defining various dimensions in focus. In sum, it could be argued that verbal responses have a higher ecological and face validity compared with rating scales.

There has been a lack of methods for quantifying language (i.e., mapping words to numbers) to capture self-reported psychological constructs. However, computational methods have been introduced in social sciences as a method for quantifying language (e.g., Kern et al., 2016; Neuman & Cohen, 2014; Park et al., 2014; Pennebaker, Mehl, & Niederhoffer, 2003). A commonly used automated text analysis approach within psychology is the word frequency procedure by Pennebaker, Francis, and Booth (2001) referred to as Linguistic Inquiry and Word Count (LIWC). Word frequency approaches are based on hand-coded dictionaries, where human judges have arranged words into different categories, such as *pronouns*, *articles*, *positive emotions*, *negative emotions*, *friends*, and so forth (Pennebaker, Boyd, Jordan, & Blackburn, 2015). Thus, the results from a LIWC analysis reveal the percentages of words in a text categorized in accordance with these predefined categories. This technique has been successful in examining *how* individuals *use* the language (e.g., their writing style), but less successful in examining the content of *what* is being said (Pennebaker, Mehl, & Niederhoffer, 2003). Pennebaker et al., (2003) point out that content-based dictionaries encounter problems dealing with all the possible topics people discuss, thus leading to poorer results due to difficulties in terms of categorizing all of these possible topics.

LIWC also suffers from other drawbacks. Categorizations of words as either belonging or not belonging to a category does not reflect the fact that words are more or less prototypical of a category. For example, words such as *neutral*, *concerned*, and *depressed* differ in negative emotional tone, whereas LIWC would require each word to either belong or not belong to a negative emotion category. LIWC also fails to capture complex nuances between words. For example, in LIWC2015 (Pennebaker et al., 2015) *love* and *nice* both belong to the positive emotion category, which fails to represent differences in, for example, their valence and arousal. Hence, this binary method is limited in terms of capturing nuances in language and complex interrelationships between words. A more precise measure should acknowledge the degree of semantic similarity between words. Pennebaker et al. (2003) conclude that "the most promising content or theme-based approaches to text analysis involve word pattern analyses such as LSA [Latent Semantic Analysis]" (p. 571). LSA allows researchers to automatically map numbers to words within a language.

Sikström has recently developed a web-based program called Semantic Excel (www.semanticexcel.com), which performs different natural language processing analyses (Garcia & Sikström, 2013; Gustafsson Sendén, Sikström, & Lindholm, 2015; Kjell, Daukantaitė, Hefferon, & Sikström, 2015; Roll et al., 2012). To quantify texts, Semantic Excel is currently based on a method similar to that of LSA, which is both a theory and a method for acquiring and representing the meaning of words (Landauer, 1999; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998; Landauer, McNamara, Dennis, & Kintsch, 2007).

LSA assumes that the contexts in which a specific word appears convey information about the meaning of that word. Hence, the first step in LSA involves representing text as a matrix of word co-occurrence counts, where the rows represent words and the columns represent a text passage or some other kind of context (e.g., Landauer et al., 1998). Singular value decomposition (Golub & Kahan, 1965) is applied to this matrix to reduce the dimensionality, and this high-dimensional structure is referred to as a semantic space. Ultimately, each word is represented by an ordered set of numbers (a vector), referred to as a semantic representation. These numbers may be seen as coordinates describing how the words relate to each other in a high-dimensional semantic space. The

closer the "coordinates" of two words are positioned within the space, the more similar they are in terms of meaning.

LSA has performed well in several text processing tasks, such as determining whether the contents of various texts have similar meaning (Foltz, Kintsch, & Landauer, 1998). Further, McNamara (2011) points out that LSA has inspired the development of new methods for producing semantic representations, such as Correlated Occurrence Analog to Lexical Semantics (COALS; Rohde, Gonnerman, & Plaut, 2005). We believe that semantic representations from other similar methods would also work for our proposed method using semantic measures. However, importantly the LSA approach has recently been successfully applied within psychology for predicting psychological constructs from texts (Arvidsson, Sikström, & Werbart, 2011; Garcia & Sikström, 2013; Karlsson, Sikström, & Willander, 2013; Kjell et al., 2015; Kwantes, Derbentseva, Lam, Vartanian, & Marmurek, 2016; Roll et al., 2012). Nevertheless, these predictions are typically performed on free texts written for very different purposes, such as status updates on Facebook, autobiographical memories, description of parents, and essays answering hypothetical scenarios. Hence, they are not collected with the objective of efficiently measuring a specific construct. They are thus not tailored for complementing rating scales, which probably is the most dominant approach to collect self-report data in behavioral science to date.

To the best of our knowledge, natural language processing has *not* been applied as a method for complementing and extending rating scales. Compared with LIWC's binary categorizations of words, LSA provides a finer and continues measure of semantic similarity, which reflects the nuances between semantic concepts. We believe that LSA's multidimensional representation of the meaning of words is particularly suitable for measuring psychological constructs. Further, we argue that the multidimensional information included in this natural language processing method may complement and extend the one-dimensional response formats of rating scales. This is why we aim to adapt and further develop these LSA-based computational methods in order to create a semantic measures method capable of both measuring and describing a targeted construct based on open-ended responses to a semantic question. This method has the potential of statistically capturing how individuals naturally answer questions about any type of psychological construct.

Methodological and Statistical Approach

The various semantic representations and analyses used in this article are described next, and descriptions of key terms relating to the developed method are presented in Table 1. The first part describes the semantic space and its semantic representations. This includes how semantic representations are added to describe several words and how artifacts relating to frequently occurring words are controlled for. The second part describes how the semantic representations are used for different analytical procedures in the forthcoming studies. This includes testing the relationship between words and a numerical variable (i.e., semantic-numeric correlations), predicting properties such as the valence of words/text (i.e., semantic predictions), measuring the semantic similarity between two words/texts (i.e., semantic similarities) and testing whether two sets of words/texts statistically differ (i.e., semantic *t* tests). Semantic Excel was used for all included studies, as it is capable of both producing and analyzing semantic representations.

The Semantic Space and Its Semantic Representations

The semantic space. An approach similar to LSA as implemented by Landauer and Dumais (1997) is employed to produce the semantic space and its semantic representations. Creating high-quality semantic representations requires a very large dataset, much larger than the data collected within the current studies. Therefore, a semantic space was created to function as a "model" for the smaller data sets generated in the current studies. Whereas some researchers produce domain-specific semantic spaces (e.g., if diaries are studied, the semantic space is based on a large amount of text from other diaries), we instead use a general semantic space. Although this might to some extent decrease the domain-specific semantic precision of semantic representations, it does make different studies more comparable with each other, while also simplifying analyses.

The current semantic space was originally created using a massive amount of text data (1.7×10^9 words) summarized in the

Table 1
Brief Description of Key Terms Relating to Semantic Measures

Term	Description
Semantic measures	Umbrella term for measures based on semantic representations.
Semantic question/item	A question/item developed to produce responses appropriate for analyses of semantic representations.
Semantic word/text response	The response to a semantic question (e.g., descriptive words).
Word-norm	A collection of words representing a particular understanding of a construct; a *high* word-norm refers to the construct under investigation (e.g., happy) and optionally a *low* word-norm refers to the construct's opposite meaning (e.g., not at all happy).
Semantic space	A matrix (here based on LSA) in which words (in rows) are described on several dimensions (in columns) that represent how all words relate to each other.
Semantic representation	The vector (i.e., an ordered set of numerical values) that words are assigned from the semantic space.
Semantic similarity score	Is the value specifying the semantic similarity between two words/texts, derived by calculating the cosine of the angle between the semantic representations of each word/text.
Unipolar scale	The semantic similarity between semantic responses and a high (or low) word norm.
Bipolar scale	The semantic similarity of the high norm minus the semantic similarity of the low norm.
Semantic-numeric correlation	The relationship between the semantic representations of words and a numeric variable such as a rating scale.
Semantic prediction	The predicted/estimated property of a word/text such as valence.
Semantic *t*-test	The statistical test of whether two sets of words/texts differ in meaning.

English (Version 20120701) Google N-gram database (https://books.google.com/ngrams). LSA typically uses document contexts (Landauer & Dumais, 1997), whereas COALS uses a (ramped) "window" of four words (i.e., the four closest words on both sides of a target word, where the closest words have the highest weight; Rohde et al., 2005). Rohde et al. (2005) found that smaller text corpora require larger windows, whereas for larger text corpora smaller window sizes are adequate without sacrificing performance. Because our corpus may be considered very large, we use 5-gram rather than documents. Using 5-gram contexts from the database, a co-occurrence (word by word) matrix was set up, where the rows contained the 120,000 most common words in the n-gram database and the columns consisted of the 10,000 most common words in the n-gram database.[1] Hence, each cell in the co-occurrence matrix denoted the frequency at which the words in the associated row/column co-occur within the 5-gram.

To increase/decrease the importance of infrequent and frequent words, log-frequency was used; meaning that the cells were normalized by computing the natural logarithm plus one (for different weighting functions, see for instance Nakov, Popova, & Mateev, 2001). Singular value decomposition was then used for compressing the matrix (while at the same time preserving as much information as possible). This was carried out to keep the most informative data while leaving out "noise." To identify how many dimensions to use in the space, some kind of a synonym test is typically conducted. Landauer and Dumais (1997) use a relatively short synonym test taken from a Test of English as a Foreign Language (TOEFL); however, we applied a more extensive test. This test analyzed how closely synonym word pairs from a thesaurus are positioned to each other in relation to the other words in the semantic space. Thus, the quality of the semantic space is measured by the rank order of the number of words positioned closer to one of the two synonym words than the two words are positioned to each other. The fewer words that are positioned closer to any of the synonym words than the synonyms themselves, the better quality is attributed to the semantic space. Testing the sequence of the powers of 2 (i.e., 1, 2, 4, . . . 256, 512, 1024) dimensions, we found that the best test score was achieved using 512 dimensions, which was subsequently considered the optimal number of dimensions. This semantic space has been demonstrated to be valid in previous research (e.g., Kjell et al., 2015).

The semantic representation of participant responses. By adding the semantic representations of single words (while normalizing the length of the vector to one), one may capture and summarize the meaning of several words and paragraphs. Hence, the words generated by participants were assigned their semantic representations from the semantic space; and then all the words' representations for an answer were being summed up and normalized to the length of one. However, the word/text responses generated by the participants were first cleaned using a manual procedure assisted with the spelling tool in Microsoft Words, so that words were spelled according to American spelling. Misspelled words were corrected when the meaning was clear or were otherwise ignored. Instances of successively repeated words or where participants had written "N/A" or the like were excluded. Answers including more than one word in response boxes only requiring one descriptive word were also excluded.

Controlling for artifacts relating to frequently occurring words. When aggregating the words to a semantic representation, a normalization is conducted to correct for artifacts related to word frequency. This is achieved by first calculating a frequency-weighted (taken from Google N-gram) average of all semantic representations in the space (x_{mean}), so that the weighting is proportional to how frequently the words occur in Google N-gram. This representation is then subtracted prior to aggregating each word and then added to the final value (i.e., Normalization $(\sum_{i=1}^{N}(x_i - x_{mean})) + x_{mean}$), where N is the number of words in the text used for creating the semantic representation, and Normalization is a function normalizing the length of the resulting vector to one.

Using the Semantic Representations in Analyses

Semantic-numeric correlations. The semantic representations can be used for analyzing the relationship between semantic responses and a numerical variable (e.g., numerical rating scale scores). This is achieved by first translating the semantic responses into corresponding semantic representations (as described above), followed by predicting the corresponding numeric rating scales on the basis of these representations by means of multiple linear regression analyses; that is, $y = \beta_0 + \beta_1 * x_1 \ldots \beta_m * x_m + \varepsilon$, where y is the to-be predicted numerical value, β_0 is the intercept, x_1 through x_m are the predictors (i.e., the values from the m number of semantic representations; x_1 = the first dimension of the semantic representations, x_2 the second dimension and so on) and β_1 through β_m are the coefficients defining the relationship between the semantic dimensions and the rating scale scores. When the predicted variable is categorical, multinomial logistic regression is used.

To avoid overfitting the model, not all semantic dimensions are used. The number of semantic dimensions to be included in the analysis is here optimized by selecting the first dimensions that best predict the actual score as evaluated using a leave-10%-out cross-validation procedure. More specifically, the optimization involves testing different numbers of dimensions, starting with the first dimension(s), which carries most of the information and adding more until all dimensions have been tested. The set of dimensions that best predict the outcome variable is finally used. The sequence used for adding more semantic dimensions aims to initially increase a few dimensions each time and then gradually increase by larger numbers. In practice, this was simply achieved by adding 1, then multiplying by 1.3 and finally rounding to the nearest integer (e.g., 1, 3, 5, 8, where the next number of dimensions to be tested are the first 12; in other words ([8 + 1] * 1.3). In previous research, we have found this sequence to be valid and computationally efficient (e.g., see Kjell et al., 2015; Sarwar, Sikström, Allwood, & Innes-Ker, 2015). Subsequently, by using leave-10%-out cross-validation, the validity of the created seman-

[1] Although some remove the most common words such as *she*, *he*, *the*, and *a*, we keep them, because some of these words may be of interest (e.g., see Gustafsson Sendén, Lindholm, & Sikström, 2014) and valid results can be achieved even when keeping them.

tic predicted scales may be tested by assessing its correlation to the numeric variable.

Semantic predictions. To further understand the relationship between words and numerical variables, it is often helpful to estimate various semantic-psychological dimensions of words such as valence. The semantic representations may be used for estimating these dimensions based on independently trained models. This approach uses previously trained semantic models and applies them to new data. In the current studies, we use the Affective Norms for English Words (ANEW; Bradley & Lang, 1999) to create a semantic trained model for valence. ANEW is a preexisting comprehensive list of words where participants have rated the words according to several dimensions, such as valence (ranging from unpleasant to pleasant). This enabled us to train each word in the word list to its specific valence rating with a semantic-numeric correlation of $r = .72$ ($p < .001$, $N = 1{,}031$). This trained model of valence is then applied to participants' answers to the semantic questions in order to create a semantic predicted scale of ANEW valence.

Semantic similarity scales. To measure the level of similarity between a semantic response and a psychological construct, we compute the semantic similarity between semantic responses and word norms, which describes the endpoints of the psychological dimensions. For example, a high semantic similarity between the answer to a semantic question regarding harmony in life and the word norm for harmony suggests a high level of harmony in life (see Figure 1). These word norms are developed by an independent sample of participants generating words describing their view of the constructs being studied (high norms, e.g., "harmony in life") and their opposite meaning (low norms, e.g., "disharmony in life"). The semantic representations of the word norms are generated from participants' replies to word norm questions, where the semantic representations from all the generated words are summed up and the resulting vector is normalized to the length of one.

The semantic similarity between two vectors is calculated as the cosines of the angle. Because the vectors have been normalized to the length of one, this calculation is the same as the *dot product* (i.e., $\sum_{i=1}^{m} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_m b_m$, where a and b refers to the two semantic representations, Σ refers to summation and m is the number of semantic dimensions). These semantic similarity scales between the semantic representations of semantic responses and word norms are either *unipolar* (i.e., similarity to the high norm of the to-be-measured construct) or *bipolar* (i.e., high minus low similarity values).

Semantic *t*-test. The difference between two sets of texts may be tested using the semantic representations (cf. Arvidsson et al., 2011). This is achieved by first creating a semantic representation reflecting the semantic difference between the two sets of texts; we refer to this vector as a *semantic difference representation*. The semantic similarity is then measured between this semantic difference representation and each individual semantic response. Finally, these semantic similarity values of the two sets of texts may be compared by means of a t test. However, to avoid biasing the results, these steps also include a leave-10%-out procedure.

This is specifically carried out by leaving-out 10% of the responses before adding the semantic representations for each of the two sets of texts to be compared. Then one of the two semantic representations is subtracted from the other to create the semantic difference representation. The semantic similarity is computed between the semantic difference representation and the semantic representations of each semantic response initially left out in the leave-10%-out procedure. The leave-10%-out procedure is repeated until all responses have been used to produce a semantic similarity score. Finally, the difference in semantic similarity between the two sets of texts is tested using a standard t test in order to attain a p value and an effect size.

In two different methodological paradigms, we empirically examine the validity and reliability of semantic measures in relation to traditional numeric rating scales. First, we develop semantic measures to describe external stimuli and examine whether these measures more accurately categorize facial expressions and yield higher interrater reliability compared with numerical rating scales. Second, we develop semantic measures for subjective states and examine the validity and reliability of subjective reports related to mental health. In all analyses, alpha was set at .05.

Reports Regarding External Stimuli

Studies 1 and 2 focused on reports regarding external stimuli, where participants made judgments based on facial expressions, and reports regarding different aspects of the stimuli. The term *external* is here used to emphasize that all participants describe (their subjective interpretation of) identical facial expressions rather than, for example, describing their subjective states (which

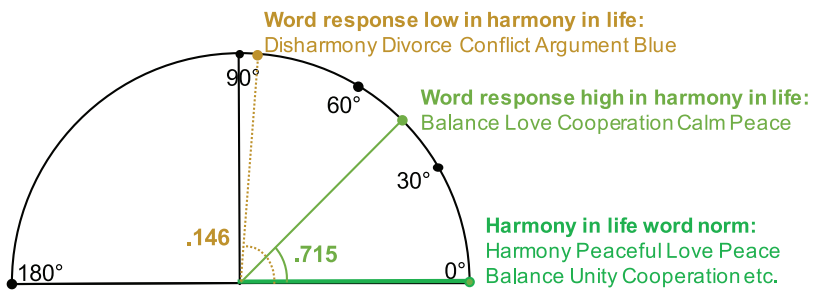

Figure 1. A conceptual illustration of the semantic similarities between word responses and a word norm. The word response high in harmony in life is positioned closer to the harmony in life word norm than the response low in harmony in life. Decimal numbers represent cosine, where the word response high in harmony in life yields a higher cosine (i.e., semantic similarity) than the word response low in harmony in life. See the online article for the color version of this figure.

is the focus in Studies 3–9). We tested whether the semantic measures were capable of categorizing and describing facial expressions from a validated picture database, including happy, sad, and contemptuous (Langner et al., 2010). The pictures of facial expressions were selected from a validation study in which human models were trained and coached by specialists to reliably express highly *prototypical* and *recognizable* facial expressions (Langner et al., 2010). Hence, these validated facial expressions were seen as the gold standard in relation to categorizing the responses as correct or not. The semantic questions required participants to describe the expressions using three descriptive words. This was compared with the methodological design validating the database (Langner et al., 2010), where expressions were categorized using checkboxes or rated using 5-point scales covering the following dimensions: degree of happiness, sadness and contempt, as well as valence, arousal, intensity, clarity, and genuineness.

The pictures were selected based on the highest interrater agreement of the targeted facial expression (Langner et al., 2010). Our studies included pictures depicting neutral, happy, sad, and contemptuous expressions. Neutral expressions were used for ensuring that the design remained the same as the original validation study, in which participants first evaluated the attractiveness of each model. Happy and sad were selected to align with the forthcoming part on subjective reports regarding mental health. Contemptuous was selected as it has been shown in previous studies that this expression is difficult to label, and thus had the lowest interrater reliability (Langner et al., 2010). This difficulty demonstrates a suitable challenge for the semantic questions approach, where it may potentially provide information on how participants perceive and describe this expression. Semantic questions enable participants to openly describe their perception of the face, rather than asking participants to tick a checkbox (Study 1) or use a labeled rating scale (Study 2). Hence, semantic questions are proposed to yield a higher level of accuracy when categorizing pictures as well as higher interrater reliability.

Method

Participants

Participants were recruited using Mechanical Turk (www.mturk.com), which is an online platform that enables offering payment to individuals to partake in studies. Mechanical Turk as a means of collecting research data within psychology has demonstrated results similar to more conventional methods, as well as good generalizability (Buhrmester, Kwang, & Gosling, 2011). The samples for the studies concerning external stimuli are described in Table 2.

Instruments and Material

Pictures of facial expressions from the Radboud Faces Database (Langner et al., 2010) were used. Study 1 and 2 included pictures of six face models displaying four different facial expressions, including happy, sad, contemptuous, and neutral.

Rating scales for external stimuli included the instruction to: "Please look at the picture below. Answer the questions about how *you* interpret the expression." As in the validation study (Langner et al., 2010), Study 1 included nine checkbox alternatives describing different expressions (i.e., "happy," "sad," "contemptuous," "angry," "disgusted," "fearful," "surprised," "neutral," and "other"); as well as 5-point scales for the intensity ("weak" to "strong"), the clarity ("unclear" to "clear"), the genuineness ("fake" to "genuine"), and the valence ("negative" to "positive") of the expressions. In Study 2, the nine alternatives were replaced by three rating scales concerning to what extent the three expressions

Table 2
Information About Participants Within Each Study for Objective Stimuli

Study: Condition	*N* (Excluded due to control questions)[1]	Age Min-Max; Mean (*SD*) years	Gender	Nationality	Mean time (*SD*) in minutes and seconds	Payment in US$
Study 1: Rating scales condition	148	18–68; 34.45 (13.04)	F = 56.8% M = 42.6% O = .7%	US = 87.2% IN = 10.8% O = 2%	9.32 (5.47)	.50
Study 1: Semantic questions condition	119	20–65; 35.70 (11.28)	F = 58.8% M = 40.3% O = .8%	US = 89.1% IN = 6.7% O = 4.2%	14.38 (8.31)	.50
Study 2: Rating scales condition	183 (7)	19–74; 34.25 (11.67)	F = 54.7% M = 44.8% O = .6%	US = 80.7% IN = 17.1% O = 2.2%	11.23 (8.08)	.50
Study 2: Semantic questions condition	134 (5)	19–67; 34.03 (11.57)	F = 63.2% M = 36.1 O = .8	US = 85% IN = 12.8% O = 2.3%	13.44 (8.20)	.50
Word-norms						
Face-norms: Expressions	107 (3)	19–61; 32.22 (9.54)	F = 51.4% M = 47.7% O = .9%	US = 95.3% IN = 3.7% O = .9%	6.56 (4.59)	.40
Face-norms: Dimensions	107 (3)	20–64; 33.97 (10.20)	F = 62.6% M = 37.4%	US = 81.3% IN = 15.9% O = 2.8%	19.22 (14.02)	.40

Note. F = Female; M = Male; O = Other; US = United States of America; IN = India.
[1] In Study 1 there were no control questions.

"happy," "sad," and "contemptuous" were expressed in the pictures (1 = *not at all* to 5 = *very much*). This condition also included a rating scale for degree of arousal (1 = *low* to 5 = *high*). With regard to the neutral pictures, only the rating scale for attractiveness was presented.

Semantic items for external stimuli first included the following general instructions: "Please look at the pictures. Answer the question about how *you* interpret the expressions. Please answer with one descriptive word in each box." The instructions were followed by the following semantic item: "Describe the expression in the picture with three descriptive words." For the neutral pictures in Study 1, the attractiveness question was phrased: "How attractive do you think the person in the picture is? Please answer with one descriptive word in each box." In Study 2, the first part was further clarified to read: "How unattractive or attractive do you think the person in the picture is?" Three boxes were presented underneath each question.

Word norms for external stimuli were generated by asking participants to imagine an expression. This was carried out to attain a general norm, which is not dependent on specific pictures depicting an expression. The following instructions were used: "Please imagine that you should describe the expression of a face in a picture. Write five words that best describe a facial expression of happy and five words that best describe a facial expression of not at all happy. Please only write one word in each box." These instructions were adapted for "sad" and "contemptuous," as well as to cover the rating scales used in the validation study of the pictures, including: valence (i.e., negative vs. positive), arousal (i.e., low vs. high), intensity (i.e., weak vs. strong), clarity (i.e., unclear vs. clear), genuineness (i.e., fake vs. genuine), and attractiveness (i.e., unattractive vs. attractive). The targeted word (e.g., happy for the happy norm) was added to the norm with a frequency of one word more than the most frequently generated word by the participants (i.e., f(max) + 1; see discussion in the online supplementary material [OSM]).

Procedure

All studies were carried out using the following general structure. First, participants were informed about the survey, confidentiality, their right to withdraw at any time without giving a reason and that they could contact the research leader with any questions about the survey. They were asked to agree to the consent form and subsequently complete the survey. Last, demographic information was collected and participants were presented with debriefing information.

In both Study 1 and 2, participants evaluated the various faces from the Radboud Faces Database. The studies involved two conditions: semantic questions and rating scales. Participants were randomly assigned to one of the conditions within the survey. The semantic questions condition was the same for Study 1 and 2, in which the semantic questions and instructions were presented with each of the 24 pictures. In both studies, participants started evaluating the randomly presented neutral pictures in relation to their attractiveness, followed by the randomly presented pictures depicting the different expressions. In the rating scales condition of Study 1, the same design as in the validation study (Langner et al., 2010) was used in which the nine facial expression alternatives were presented, whereas the three rating scales were presented in Study 2. The word norms relating to the reports on external stimuli were collected in two separate studies; one for the expressions and another for the dimensions. The word norm questions were presented in random order.

Results

The detailed results for Study 1 are presented in the OSM. Descriptive data of the number of words generated by participants in the semantic questions is presented in Table S1 in OSM. In Study 2, the semantic questions and the rating scales conditions both involved a one third chance of being correct through random categorization. The semantic responses produced 83.1% correct categorizations of facial expressions when using semantic predicted scales. This was achieved by training the semantic representations to the expressed facial expression using semantic predicted scales based on multinomial logistic regressions where participants were used as the grouping variable. When grouping the training to pictures, the overall level of correctness reached 78.8%. In comparison, the rating scales responses produced an overall correct categorization (i.e., the targeted expression receiving the highest rating score) of 74.2%. Hence, grouping according to participants, $\chi^2(1) = 62.95, p < .001$ or pictures, $\chi^2(1) = 15.75$, $p < .001$ yields significantly higher levels of overall correct categorizations compared with using rating scales.

It is also possible to categorize the facial expressions using word norms. The semantic similarity scores between the responses to each picture and the word norms was normalized to z-scores. A response was categorized as correct when the semantic similarity with the word norm representing the facial expression depicted in the picture was the highest. When using semantic similarity scales, the overall correct categorization reached 80.7% with unipolar scales (i.e., happy, sad, and contemptuous) and 80.1% with bipolar scales (i.e., subtracting "not at all happy," "not at all sad," and "not at all contemptuous" for each construct, respectively). Hence, both unipolar and bipolar scales yield significantly higher correct categorizations compared with numerical rating scales (unipolar: $\chi^2(1) = 32.19$; bipolar: $\chi^2(1) = 26.82$; $p < .001$). The significantly higher level of categorization using semantic measures supports the validity of this method.

In the rating scales condition, the overall correctness was 74.2%. Ties were categorized as incorrect, so a correct score required the rating scale score to be higher on the rating item for the targeted expression (e.g., happy) than for the two other rating scale items concerning facial expression (i.e., sad and contemptuous). This is arguably the most straightforward approach, especially as the validated database of pictures included *prototypical* facial expressions based on the Facial Action Coding System (see Langner et al., 2010). This system emphasizes the anatomy of facial expressions (e.g., Bartlett et al., 1996), where the pictures include basic emotions in which the expressions are frequently recognized across cultures (Ekman, 1992). However, a less stringent approach is to split the correctness point between the ties, so that .50 is given to answers where one of the two highest scores include the targeted expression, and .33 is given when the same rating scores have been given for all three rating scales items. Employing this approach yields an overall correct categorization of 80.7%.

However, it is important to point out that affording the rating scales condition this advantage does not make it significantly

better compared with the semantic measures. Training the semantic questions responses using participants as a grouping variable still produces significantly higher levels of correct categorizations, $\chi^2(1) = 4.95$, $p = .026$, whereas the comparisons with the other methods do not exhibit any significant differences. On the other hand, the semantic method could be improved by means of several other factors. For example, one could control for missing values within the semantic questions condition. That is, missing values could arguably be treated as 1/3 of a correct answer in order to become more comparable with the rating scale condition. The semantic categorization may be further improved upon by adjusting weights to each semantic similarity scale; however, this is outside the scope of this article.

To analyze interrater reliability, both trained and similarity scales were studied. Unipolar and bipolar scales were used for the categorization of expressions, and bipolar scales were used for the other dimensions. For the semantic trained scales of valence, intensity, clarity, genuineness, and attractiveness, the semantic representations from the semantic questions were first trained to the mean score for each dimension from the rating scales condition, and then validated using leave-10%-out cross-validation.

Semantic measures yield significantly higher interrater reliability compared with rating scales with regard to both categorizing expressions and describing the related dimensions. All trained models, except for the attractiveness model grouped according to pictures, were significant in Study 2 (Pearson's *r* ranged from .53 to .91 for the significant models [$p < .001$]; see Table S5). The interrater reliability statistics were measured by both intraclass correlation coefficient (ICC; two-way agreement) and Krippendorff's alpha. The differences between approaches were tested by bootstrapping Krippendorff's alpha (1,000 iterations), followed by significance testing the results using *t* tests between the various semantic measures and corresponding rating scale. As the bootstrapping procedure may end up computationally intensive, scores in the semantic conditions were rounded to two significant digits. Importantly, the ICC values tend to be higher using semantic measures as compared with rating scales (see Table 3). The *t* tests revealed that the semantic measures yield a significantly higher Krippendorff's alpha compared with the rating scales for all dimensions ($p < .001$), except for attractiveness, where rating scales were significantly higher ($p < .001$; Table S6). Further, applying semantic predictions of ANEW valence to the semantic responses also yields a high interrater reliability ($ICC_{(2,1)} = .767$; Krippendorff's $\alpha = .74$, CI [.72, 76]).

Analyzing word responses also enables statistically based word descriptions of the constructs being studied. Figure 2 plots words that differ statistically between sets of word responses on the x-axes and words that statistically relate to the semantic similarity scales on the y-axes. In this way, the plots describe discriminative characteristics between the various constructs in focus. And even though participants have not been primed with the psychological constructs, happy facial expressions were described as *happy* and *joyful*, sad facial expressions were described as *sad* and *unhappy* and contempt facial expressions were described as *annoyed* and *irritated.*

Discussion

The results from Study 2 suggest that the semantic measures encompass higher validity and interrater reliability compared with rating scales, except for evaluations of attractiveness; which was also supported by the results in Study 1. Even though the method used for categorizing facial expressions differed between Study 1 (using check boxes) and 2 (using rating scales), the results from the semantic conditions are to a high extent replicated between the studies. Importantly, the ability of semantic measures to correctly categorize facial expressions to a higher degree than rating scales in Study 2 demonstrates their ability to *differentiate* between measured constructs.

The clarification of the semantic question of attractiveness used in Study 2 revealed an improvement of interrater reliability from Study 1. However, the low interrater reliability might reflect that perceived attractiveness is subjective in nature and thus should not encompass high interrater reliability, as participants perceive the various models differently. That is, rating attractiveness is more subjective than categorizing prototypical facial expressions. Interrater reliability may potentially be further increased by additional clarification of the semantic question; for example, by making sure

Table 3
Krippendorff's Alpha (α) and Interclass Correlation Coefficient (ICC)

	Three descriptive words (*N*: α =133; ICC = 127)[1]							
	Trained semantic scales							
	Participants		Pictures		Bipolar similarity scales		Rating scales (*N* = 181)	
Dimensions	α [95% CI]	ICC (2,1)	α [95% CI]	ICC (2,1)	α [95% CI]	ICC (2,1)	α [95% CI]	ICC (2,1)
Expression[2]	.60 [.52, .69]	NA	.53 [.44, .62]	NA	.58 [.49, .66][3] .57 [.48, .66][4]	NA	.48 [.44, .52]	NA
Valence	.81 [.80, .82]	.840	.81 [.80, .81]	.806	.82 [.81, .82]	.835	.31 [.23, .38]	.318
Arousal	.77 [.75, .78]	.802	.77 [.75, .78]	.806	.72 [.71, .72]	.734	.13 [.03, .21]	.137
Intensity	.64 [.61, .66]	.694	.62 [.60, .64]	.667	.53 [.51, .54]	.553	.24 [.15, .32]	.249
Clarity	.66 [.64, .67]	.708	.64 [.62, .65]	.683	.67 [.67, .68]	.694	.24 [.16, .33]	.255
Genuineness	.77 [.75, .79]	.817	.77 [.75, .79]	.820	.68 [.67, .69]	.706	.17 [.08, .25]	.178
Attractiveness	21 [.17, .24]	.317	.01 [−.09, .10]	.002	.21 [.19, .22]	.255	.31 [.23, .38]	.354

[1] ICC is sensitive to missing values, which is why participants are removed so that these computations are based on 127 participants in the semantic question condition. [2] Expressions are based on nominal data; in the semantic questions condition there are three categories; meanwhile in the rating scale condition there are seven categories including the various combinations of ties. [3] Unipolar. [4] Bipolar.

Figure 2. *a–c.* Words that significantly describe one facial expression while covarying for the other two facial expressions. On the *x*-axis, words are plotted according to the χ-values from a chi-square test, with Bonferroni correction for multiple comparisons (Bonf. = Bonferroni line where $\chi = 4.26$, .05 indicates the uncorrected *p* value line, where $\chi = 1.96$). On the *y*-axis, words are plotted according to point-biserial correlations ($r = .09$ at the Bonferroni line, and $r = .04$ at the .05 uncorrected *p* value line). More frequent words are plotted with a larger font size with fixed lower and upper limits. The x-axes represent χ-values associated to (a) words generated to the sad and contemptuous pictures (black/left) versus words generated to happy pictures (green/right); in (b) happy and contemptuous words (black/left) versus sad words (green/right); and in (c) happy and sad words (black/left) versus contemptuous words (green/right). The y-axes show significant words related to semantic similarity (SS) scales; cov. = covariate. $N = 133$. See the online article for the color version of this figure.

that all participants interpret attractiveness in a similar way. Participants may, for example, be instructed to only evaluate attractiveness rather than describing other facial characteristics they believe to be related to attractiveness. Instructions could further aim to limit the risk that participants evaluate potentially different kinds of attractiveness, such as beauty, cuteness or sexual attractiveness.

Semantic measures also have an advantage over rating scales in terms of not priming participants with concepts required for defining the numerical scales. That is, despite the fact that participants were not primed with the targeted expressions, happy expressions were most frequently described using the word *happy* and sad using the word *sad*. Further, previous research reports low interrater agreement with regard to contempt and has argued that this is the result of issues with the label of the expression rating scale rather than the expression itself (Langner et al., 2010). In contrast, the semantic questions approach lets participants generate a description of the expression, including *annoyed* and *irritated*.

This means that whereas rating scales only provide a number in relation to the question/item, semantic measures enhance the understanding of what was measured by providing a description of it. In sum, the results demonstrate that the semantic measures may be used for measuring, differentiating, and describing psychological dimensions of facial expressions.

Reports Regarding Subjective States

Next, we develop semantic measures for two psychological constructs pertaining to well-being: harmony in life and satisfaction with life, as well as two psychological constructs relating to mental health problems: depression and worry. These constructs are theoretically different, but are frequently found to be highly correlated when measured by rating scales. In these studies, the answers have no relation to external stimuli identical to all participants, which was the case in the previous facial expression studies. Instead, we analyze how the self-reported subjective responses relate to each other across psychological constructs and response formats (i.e., semantic questions vs. rating scales).

Satisfaction with life, here measured with the numerical rating scale the Satisfaction with Life Scale (SWLS; Diener et al., 1985), focuses on evaluations based on comparing actual with expected and desired life circumstances. Harmony in life is here measured using the numerical rating scale the Harmony in Life Scale (HILS; Kjell et al., 2015). Li (2008) points out that: "Harmony is by its very nature relational. It is through mutual support and mutual dependence that things flourish" (p. 427). Hence, the HILS focuses on psychological balance and interconnectedness with those aspects considered important for the respondent. Kjell et al. (2015) found that harmony in life and satisfaction with life differ significantly in terms of how participants perceive their pursuit of each construct. The pursuit of harmony in life is significantly related to words linked to interdependence and being at peace such as *peace*, *balance*, *calm*, *unity*, and *love*, whereas the pursuit of satisfaction with life is significantly related to words linked to independence and achievements, such as *money*, *work*, *career*, *achievement*, and *fulfillment*. Hence, these (and associated) words are proposed to make up the positive endpoints representing a high degree of harmony in life versus satisfaction with life, respectively. Further, the words generated in relation to each construct significantly differed with a medium effect size (Cohen's $d = .72$) in a semantic t test (Kjell et al., 2015). This clear semantic difference between the constructs suggests that semantic measures might provide a clear differentiation between the two constructs.

Depression and worry/anxiety share a common mood factor of negative affectivity, whereas low positive affectivity is typical for depression alone (Axelson & Birmaher, 2001; Brown, Chorpita, & Barlow, 1998; Watson, Clark, & Carey, 1988). Clinical depression is characterized by symptoms such as a lack of interest or pleasure in doing things, fatigue, feelings of hopelessness and sadness (American Psychiatric Association, 2013). Depression is often measured by the nine-item Patient Health Questionnaire (PHQ-9), which targets the *DSM–IV* diagnostic criteria (Kroenke, Spitzer, & Williams, 2001). We anticipate that the semantic responses of participants in relation to depression correspond to these criteria (including words such as sad, tired, disinterested, etc.). Excessive and uncontrollable worry, on the other hand, is recognized by symptoms such as restlessness, being on edge and irritability (American Psychiatric Association, 2013). Worry is often assessed with the seven-item Generalized Anxiety Disorder Scale (GAD-7; Spitzer, Kroenke, Williams, & Löwe, 2006), which is linked to these symptoms. Thus, semantic responses to worry are anticipated to relate to these symptoms (including words such as tense, nervous, anxious, etc.).

Numerical rating scales targeting depression and worry/anxiety tend to correlate strongly with each other (Muntingh et al., 2011; Spitzer et al., 2006). This may be seen as a measurement problem associated with numerical rating scales, as correctly identifying and differentiating between the two constructs becomes difficult. Some argue that this significant overlap is due to a frequent co-occurrence between these conditions (Kessler, Chiu, Demler, Merikangas, & Walters, 2005), whereas others argue that problems differentiating these have considerable implications in terms of treatment (Brown, 2007; Wittchen et al., 2002). However, considering the conceptual and criteria-based differences between these two constructs, semantic measures are proposed to be able to differentiate between these constructs more clearly than rating scales.

The Overall Rationale of the Studies Concerning Subjective States

Studies 3–9 focused on reports regarding subjective states, where participants answered semantic questions by generating descriptive words or texts. Participants then answered numerical rating scales corresponding to the studied constructs. The semantic questions were presented first so that the items in the rating scales would not influence the generation of words and texts by participants. Because we are developing a new approach for measuring subjective states, we have carried out seven studies, which in a controlled and iterative manner enabled us to examine potential strengths and weaknesses associated with the method, as well as the replicability of the main findings (which is important considering the replicability concerns within psychological research; e.g., Open Science Collaboration, 2015).

The aim of Study 3 was to pilot the semantic questions of satisfaction with life and harmony in life and study their relationships with the corresponding numerical rating scales. The aim of Study 4 was to examine the semantic questions in a larger sample than the one used in Study 3. The aim of Study 5 involved examining how the semantic measures of harmony and satisfaction relate to rating scales of depression, anxiety and stress. Studies 3–5 included both descriptive words and text responses. The aim of Study 6 was to test shorter instructions for the semantic questions. The aim of Study 7 was to examine the effects of requiring fewer word responses, where participants only had to answer the semantic questions using one, three, or five words rather than 10. The aim of Study 8 was to develop semantic questions for depression and worry. The aims of Study 9 involved examining the test–retest reliability of the semantic measures for harmony and satisfaction, and at Time 2 (T2) examine the interrelationship between the semantic measures for all four constructs.

Method

Participants

Participants were recruited using Mechanical Turk. The samples for the studies concerning subjective states are described in Table 4.

Measures

Semantic questions concerning subjective states were developed for both using descriptive words and text responses. Pilot studies from our lab have shown that a high validity of the semantic measures requires a semantic question with instructions to be posed in a clear manner. Hence, the following guidelines were applied: Participants should in detail describe their state or perception of the to-be-answered question, for example their own life satisfaction, rather than describing their view of life satisfaction as a general concept. To get the best effect, the question and its accompanying instructions should stress the strength and frequency of words related to the construct (or lack of it). To receive a consistent response mode among participants, the instructions

Table 4
Information About Participants Within Each Study for Subjective States

Study: Condition	*N* (excluded due to control questions)	Age Min-Max; Mean (*SD*) years	Gender	Nationality	Mean time (*SD*) in minutes and seconds	Payment in US$
Study 3	92 (13)	18–64; 33.36 (10.93)	F = 51.1% M = 48.9%	US = 73.9% IN = 23.9% O = 2.2%	14.18 (8.23)	.30
Study 4	303 (24)	18–74; 34.87 (12.17)	F = 55.4% M = 44.6%	US = 95% IN = 2.6% O = 2.3%	20.06 (10.20)	.50
Study 5	296 (19)	18–74; 36.40 (13.45)	F = 60.1% M = 39.9%	US = 86.1% IN = 10.5% O = 3.4%	17.08 (21.10)	.30
Study 6	193 (9)	18–64; 35.88 (10.97)	F = 51.8% M = 48.2%	US = 78.8% IN = 18.1% O = 3.1%	10.27 (7.07)	.80
Study 7: 1 word	361 (20)	18–72; 30.80 (10.02)	F = 43.8% M = 56.0% O = .3%	US = 91.4% IN = 6.4% O = 2.2%	3.07 (2.50)	.20
Study 7: 3 words	350 (18)	18–65; 31.61 (9.65)	F = 48.6% M = 50.9% O = .6%	US = 95.4% IN = 2.6% O = 2.0%	3.35 (2.23)	.20
Study 7: 5 words	257 (19)	18–63; 30.53 (9.24)	F = 44.4% M = 55.6%	US = 94.2% IN = 3.5% O = 2.3%	4.34 (4.47)	.20
Study 8	399 (36)	18–69; 34.45 (11.45)	F = 51.0% M = 48.7% O = .3%	US = 86.2% IN = 10.6% O = 3.3%	9.44 (5.42)	.50
Study 9: T1	854 (42)	18–64; 32.76 (10.09)	F = 50.9% M = 49.0%	US = 93.5% IN = 4.1% O = 2.1%	5.53 (3.52)	.50
Study 9: T2	477 (42)	18–63; 34.11 (10.47)	F = 54.5% M = 45.5%	US = 93.9% IN = 4.4%% O = 1.7%	16.32 (9.32)	1.00
Word-norms[1]						
Harmony	120	18–51; 29.43 (7.89)	F = 40.8% M = 59.2%	US = 95.8% IN = 1.7% O = 2.5%	2.46 (2.16)	.10
Disharmony	96	18–59; 29.75 (8.49)	F = 43.8% M = 56.3%	US = 93.8% IN = 5.2% O = 1.0%	2.36 (1.28)	.10
Satisfaction	93	19–60; 29.87 (9.12)	F = 31.2% M = 68.8%	US = 98.9% O = 1.1%	2.15 (1.07)	.10
Dissatisfaction	84	18–74; 33.14 (13.35)	F = 44% M = 56%	US = 91.7% IN = 4.8% O = 3.6%	2.44 (1.41)	.10
Worried	104	18–65; 28.73 (8.80)	F = 52.9% M = 46.2% O = 1.0%	US = 93.3% IN = 4.8% O = 1.9%	2.14 (1.46)	.10
Depressed	110	18–65; 30.57 (9.64)	F = 45.5% M = 53.6% O = 0.9%	US = 89.1% IN = 7.3% O = 3.6%	1.59 (1.52)	.10

Note. F = Female; M = Male; O = Other; US = United States of America; IN = India.
[1] In Word-norms studies there were no control questions.

Table 5
Types of Semantic Questions and Numerical Rating Scales Included in Each of the Studies

Study	Semantic questions	Numerical rating scales
3	**H & S; words and text responses**	**HILS, SWLS**
4[a]	H & S; words and text responses	HILS, SWLS
5	H & S; words and text responses	HILS, SWLS, **DASS**
6	H & S; words responses; **short instructions**	HILS, SWLS
7	H & S; **1, 3, or 5** words responses	HILS, SWLS
8	**D & W**; words responses	**GAD-7, PHQ-9**
9: **T1**	H & S; words responses	HILS, SWLS
9: **T2**	H, S, D & W; words responses	HILS, SWLS, GAD-7, PHQ-9, **MC-SDS-FA**

Note. The semantic questions included the long instructions and required 10 descriptive words if nothing else is specified. Bold font highlights important (new) aspects of the study.
[a] Study 4 includes more participants than Study 3. H = Harmony in life; S = Satisfaction with life; D = Depression; W = Worry; HILS = Harmony in Life Scale; SWLS = Satisfaction with Life Scale; DASS = Depression, Anxiety, and Stress Scales the short version; GAD-7 = Generalized Anxiety Disorder Scale-7; PHQ-9 = Patient Health Questionnaire-9; MC-SDS-FA = The Marlowe-Crowne Social Desirability Scale, the short version Form A.

should also inform the participant to write *one* word in each box (except for the text responses) to discourage freestanding negations (e.g., "not happy" rather than "unhappy" or "sad"). The questions, which in combination with instructions were used for prompting open-ended responses within the studies, include: "Overall in your life, are you in harmony or not?," "Overall in your life, are you satisfied or not?," "Over the last 2 weeks, have you been worried or not?," and "Over the last 2 weeks, have you been depressed or not?" The questions are posed with different time frames (i.e., overall in your life vs. over the last 2 weeks) to reflect the timeframes prompted in the instructions/items for each respective numerical rating scale. The instructions for the semantic questions concerning harmony are presented below (these were then adapted to satisfaction with life, depression and worry).

Semantic question instructions for descriptive words responses. "Please answer the question by writing 10 descriptive words below that indicate whether you are in harmony or not. Try to weigh the strength and the number of words that describe if you are in harmony or not so that they reflect your overall personal state of harmony. For example, if you are in harmony then write more and stronger words describing this, and if you are not in harmony then write more and stronger words describing that. Write descriptive words relating to those aspects that are most important and meaningful to you. Write only one descriptive word in each box."

Short semantic question instructions for descriptive words responses. "Please answer the question by writing 10 words that describe whether you are in harmony or not. Write only one descriptive word in each box."

Semantic question instructions for text responses. "Please answer the question by writing at least a paragraph below that indicates whether you are in harmony or not. Try to weigh the strength and the number of aspects that describe if you are in harmony or not so that they reflect your overall personal state of harmony. For example, if you are in harmony then write more about aspects describing this, and if you are not in harmony then write more about aspects describing that. Write about those aspects that are most important and meaningful to you."

Word norm items for subjective states included the following instructions: "Please write 10 words that best describe your view of harmony in life. Write descriptive words relating to those aspects that are most important and meaningful to you. Write only one descriptive word in each box." The instructions were adapted to also cover "disharmony in life," "satisfaction with life," "dissatisfaction with life," "being worried," and "being depressed." (As there are no antonyms for worried and depressed, norms for these constructs were not created for the current study, although one could potentially create norms for "not worried" and "not depressed.") The targeted words were also added (i.e., f(max) + 1), as in the facial expression studies.

The Harmony in Life Scale (Kjell et al., 2015) was used in Studies 3–7 and 9. The scale includes five items (e.g., "I am in harmony"), which are answered on a scale ranging from 1 = *strongly disagree* to 7 = *strongly agree*. For the different studies, McDonald's omega (Dunn, Baguley, & Brunsden, 2014) ranged from .91–.95 and Cronbach's alpha ranged from .89–.95.

The Satisfaction with Life Scale (Diener et al., 1985) was used in Studies 3–7 and 9. It comprises five items (e.g., "I am satisfied with life") answered on the same scale as the HILS. In the various studies, McDonald's omega and Cronbach's alpha ranged from .90 to .94.

The Depression, Anxiety and Stress Scale, the short version (DASS; Sinclair et al., 2012) was used in Study 5. It includes seven items for each of the three constructs: depression (e.g., "I felt downhearted and blue"), anxiety (e.g., "I felt I was close to panic"), and stress (e.g., "I found it hard to wind down"). Participants were required to answer using a 4-point scale assessing severity/frequency of the constructs. Both McDonald's omega and Cronbach's alpha were .93 for depression and .90 for anxiety and stress.

The Generalized Anxiety Disorder Scale-7 (Spitzer et al., 2006) was used for measuring worry in Studies 8 and 9 at T2. It includes seven items answered on a scale ranging from 0 = *not at all* to 3 = *nearly every day*. It assesses how frequently the respondent has been bothered by various problems over the last two weeks (e.g., "Worrying too much about different things"). In both studies, McDonald's omega and Cronbach's alpha were .94.

The Patient Health Questionnaire-9 (Kroenke & Spitzer, 2002) was used for assessing depression in Studies 8 and 9 at T2. Its structure and response format is similar to that of GAD-7, com-

Table 6
Semantic Responses Trained to Rating Scales Scores Demonstrate Validity, Reliability, and the High Influence of Valence

Study: condition		S3 (N = 91)	S4 (N = 301)	S5 (N = 294)	S6 (N = 190)	S7: 1-w (N = 355)	S7: 3-w (N = 349)	S7: 5-w (N = 256)	S9: T1 (N = 852)	S9: T2 (N = 477)		S8 N = 399	S9: T2 N = 477
Correlation items		Pearson's r										Pearson's r	
Hw	HILS	.60***	.66***	.71***	.60***	.56***	.57***	.58***	.70***	.72***	Ww	GAD-7	.57***	.58***
	SWLS	.34***	.49***	.61***	.50***	.33***	.36***	.47***	.62***	.48***		PHQ-9	.47***	.43***
Hw: Valence as covariate	HILS	ns	ns	.23***	ns	.36***	ns	.19**	.30***	.17***	Ww: Valence as covariate	GAD-7	.14**	.12**
	SWLS	ns	ns	ns	ns	.16**	ns	.11*	.22***	ns		PHQ-9	ns	ns
Sw	SWLS	.50***	.66***	.62***	.58***	.47***	.56***	.60***	.70***	.63***	Dw	PHQ-9	.63***	.59***
	HILS	ns	.64***	.62***	.68***	.50***	.58***	.52***	.69***	.68***		GAD-7	.56***	.51***
Sw: Valence as covariate	SWLS	.25**	.12*	ns	ns	.27***	.18***	ns	.19***	.09*	Dw: Valence as covariate	PHQ-9	.20***	ns
	HILS	ns	ns	ns	.25***	.34***	ns	ns	.10**	.21***		GAD-7	ns	ns
HILS	SWLS	.80***	.84***	.88***	.84***	.83***	.83***	.84***	.84***	.83***	GAD-7	PHQ-9	.86***	.87***

Note. Pearson's r between predicted and actual values. ns = not significant; Hw = harmony words; Sw = Satisfaction words; Ww = Worry words; Dw = Depression words; HILS = Harmony in Life Scale; SWLS = Satisfaction with Life Scale; GAD-7 = Generalized Anxiety Disorder Scale-7; PHQ-9 = Patient Health Questionnaire-9; S = Study; 1-w, 3-w, and 5-w = number of words required as response to the semantic question (i.e., one, three, and five, respectively).
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 7
Correlations Between Predicted and Actual Rating Scale Scores as a Function of Number of Participants in Study 9 at T2

	Function of N participants Pearson's r			
N-ps	Hw: HILS	Sw: SWLS	Ww: GAD-7	Dw: PHQ-9
8	ns	ns	ns	ns
16	ns	ns	ns	ns
32	.40**	.20*	ns	.30*
64	.48	.43	.25**	.21
128	.63	.46	.47	.39
256	.70	.62	.51	.49
477	.72	.63	.58	.59

Note. All correlations were significant at $p < .001$ unless otherwise specified; (N = 477). ns = not significant; ps = participants; Hw = Harmony words; Sw = Satisfaction words; Ww = Worry words; Dw = Depression words; HILS = Harmony in Life Scale; SWLS = Satisfaction with Life Scale; GAD-7 = Generalized Anxiety Disorder Scale-7; PHQ-9 = Patient Health Questionnaire-9.
* $p < .05$. ** $p < .01$.

prising nine items (e.g., "Feeling down, depressed or hopeless"). In both studies, McDonald's omega and Cronbach's alpha were .93.

The Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960) the short version Form A (MC-SDS-FA; Reynolds, 1982) was used in Study 9 at T2. It encompasses 12 items (e.g., "No matter who I'm talking to, I'm always a good listener"), which require responses indicating whether the statements are personally true or false. Both McDonald's omega and Cronbach's alpha were .73.

Control questions were used in accordance to previous research (Kjell et al., 2015). These required respondents in Studies 3–9 to answer a specific alternative (e.g., "Answer 'disagree' on this question") in various places throughout the surveys. Participants who did not answer all control items correctly were excluded from the analyses, as this type of approach has been demonstrated to

Table 8
Correlations Between Predicted and Actual Rating Scale Scores as a Function of Serial Position of the Words in Study 9 at T2

	Function of word at serial position Pearson's r			
Word position	Hw: HILS	Sw: SWLS	Ww: GAD-7	Dw: PHQ-9
First	.50	.47	.48	.62
Second	.43	.47	.41	.36
Third	.42	.34	.35	.44
Fourth	.33	.29	.29	.40
Fifth	.36	.35	.35	.13
Sixth	.38	.40	.23	.24
Seventh	.22	.28	.30	.36
Eighth	.32	.25	.36	.21
Ninth	.29	.13**	.21	.30
Tenth	.37	.18	.23	.24

Note. All correlations were significant at $p < .001$ unless otherwise specified; (N = 477). Hw = Harmony words; Sw = Satisfaction words; Ww = Worry words; Dw = Depression words; HILS = Harmony in Life Scale; SWLS = Satisfaction with Life Scale; GAD-7 = Generalized Anxiety Disorder Scale-7; PHQ-9 = Patient Health Questionnaire-9.
** $p < .01$.

Table 9
Correlations Between Predicted and Actual Rating Scale Scores as a Function of the Number of Words in Study 9 at T2

	Function of *N*-first-words Pearson's r			
N words	Hw: HILS	Sw: SWLS	Ww: GAD-7	Dw: PHQ-9
1	.50	.47	.48	.62
2	.62	.56	.48	.50
3	.66	.58	.45	.57
4	.62	.58	.51	.60
5	.67	.62	.54	.59
6	.66	.62	.57	.61
7	.66	.64	.57	.63
8	.67	.64	.59	.62
9	.69	.63	.59	.61
10	.72	.63	.58	.59

Note. All correlations were significant at $p < .001$ unless otherwise specified; ($N = 477$). Hw = Harmony words; Sw = Satisfaction words; Ww = Worry words; Dw = Depression words; HILS = Harmony in Life Scale; SWLS = Satisfaction with Life Scale; GAD-7 = Generalized Anxiety Disorder Scale-7; PHQ-9 = Patient Health Questionnaire-9.

yield high statistical power and improve reliability (Oppenheimer, Meyvis, & Davidenko, 2009).

Procedures

In Studies 3–9 regarding subjective states, the semantic questions were presented first followed by the rating scales (see Table 5 for an overview of the type of semantic questions and rating scales included in these studies). Studies 3–7 started with the semantic questions concerning harmony in life and satisfaction with life in a random order. Only Studies 3–5 included both descriptive words and descriptive texts as response formats. Participants were randomly presented with either the word- or text-based items first (harmony and satisfaction were in random order). Studies 6–9 only included descriptive word-based (not text) response formats. Study 6 included the short instruction version of the descriptive word item. In Study 7, participants were asked to answer using either one, three, or five words.

In Studies 3–7 and 9, the open-ended items were followed by the HILS and the SWLS. The DASS was presented last in Study 5. Study 8 encompassed the same general structure as the one used in previous studies, but instead of asking about harmony and satisfaction, it included semantic questions concerning worry and depression, followed by the respective rating scales: the GAD-7 and the PHQ-9.

Study 9 involved a test–retest procedure. This meant that at Time 1 (T1), participants filled out the semantic questions for harmony and satisfaction followed by the HILS and the SWLS. At T2, they were asked via the message service of Mechanical Turk to partake in a follow-up study. At T2 (30.79, $SD = 2.01$, days after T1), they first filled out the questions from T1 followed by the questions from Study 8. Finally, they filled out the MC-SDS-FA.

In the development of the word norms for subjective states, participants were randomly assigned to answer one of the word norm questions and then answer the demographic battery of questions. The studies received ethical approval from the ethical committee in Lund, Sweden.

Results

Descriptive data regarding the number of words generated by participants for the semantic questions (Table S7) and the results for the descriptive text responses are presented in the OSM. Some semantic measures and rating scales were not normally distributed; in particular GAD-7 and PHQ-9 showed a positive skew (which makes sense, because these scales were originally developed for a clinical population, but have later been validated in the general population; see Löwe et al., 2008 for GAD-7 and Martin, Rief,

Table 10
Word-Norms Measure Constructs Independently From Rating Scales as Seen by Their Intercorrelations

Measure type	Numeric scales				Valence				SSS				
Measure	1	2	3	4	5	6	7	8	9	10	11	12	13
1. HILS													
2. SWLS	.83												
3. GAD-7	−.67	−.60											
4. PHQ-9	−.67	−.64	.86										
5. Hw: Valence	.74	.59	−.55	−.53									
6. Sw: Valence	.71	.67	−.51	−.52	.70								
7. Ww: Valence	.52	.48	−.61	−.52	.45	.48							
8. Dw: Valence	.59	.55	−.60	−.62	.50	.51	.72						
9. Hw: Bipolar	.65	.52	−.52	−.49	.87	.64	.42	.44					
10. Sw: Bipolar	.64	.60	−.47	−.47	.63	.89	.44	.45	.62				
11. Hw: Unipolar	.43	.34	−.38	−.36	.57	.43	.33	.33	.72	.46			
12. Sw: Unipolar	.35	.33	−.28	−.30	.32	.51	.27	.23	.36	.63	.44		
13. Ww: Unipolar	−.32	−.29	.39	.32	−.34	−.37	−.52	−.37	−.33	−.39	−.21	−.12**	
14. Dw: Unipolar	−.30	−.28	.31	.29	−.27	−.29	−.31	−.33	−.26	−.28	−.20	ns	.60

Note. Pearson's r correlations, all correlations were significant at $p < .001$ unless otherwise specified. $N = 477$. ns = not significant; HILS = Harmony in Life Scale; SWLS = Satisfaction with Life Scale; GAD-7 = Generalized Anxiety Disorder-7; PHQ-9 = Patient Health Qustionnaire-9; Hw = Harmony words; Sw = Satisfaction words; Ww = Worry words; Dw = Depression words; Valence = Semantic Predicted Valence Scale; Bipolar = Bipolar Semantic Similarity Scale; Unipolar = Unipolar Semantic Similarity Scale; SSS = Semantic Similarity Scales.
** $p < .01$.

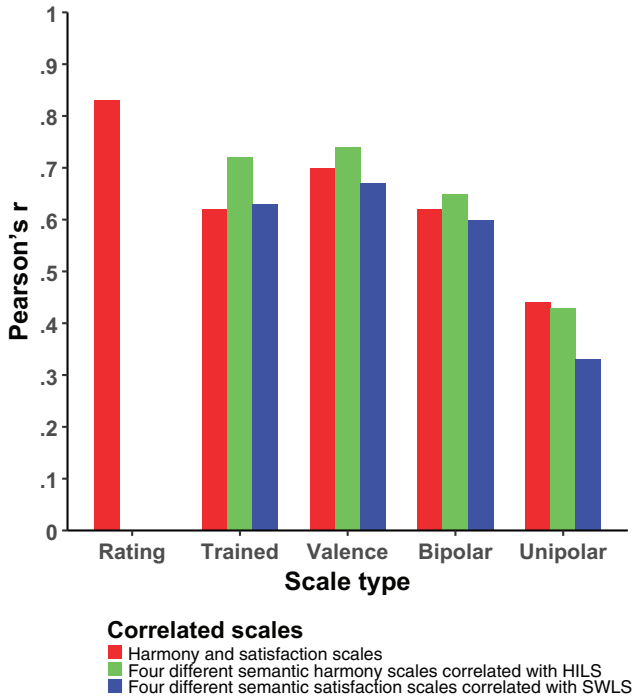

Figure 3. Pearson correlation (*y*-axis) among the various types of scales (*x*-axis) for harmony and satisfaction. The red bars show the correlation between harmony and satisfaction for one numerical measure (HILS-SWLS) and for four different semantic measures. The green bars show HILS correlated with four different semantic measures of harmony, while the blue bars similarly show corresponding correlations for SWLS and four semantic measures of satisfaction. Rating = numerical rating scales; Trained = trained predicted scales between word responses and rating scales; Valence = semantic predicted ANEW valence scales; Bipolar = bipolar semantic similarity scales; Unipolar = unipolar semantic similarity scales. See the online article for the color version of this figure.

Klaiberg, & Braehler, 2006 for PHQ-9). Consequently, it could be argued that rank-ordered statistical analyses, such as Spearman's rho, should be used in these instances. Throughout this article, however, Pearson's *r* is presented as being consistent in the studies on subjective reports and thus increase comparability across studies. Further, Pearson's *r* does not involve transforming the data into ranks and thus losing information. Importantly though, computing Spearman's rho tends to yield similar results and overall conclusions as compared with Pearson's *r* (e.g., with Spearman's rho in Study 9 at T2, there tends to be slightly smaller correlation coefficients for the well-being measures and larger coefficients for the ill-being measures).

Predicting Rating Scale Scores From Semantic Word Responses

Using semantic-numeric correlations, the predictive validity of semantic word responses was examined in relation to the rating scales. Table 6 shows that the semantic responses may be trained to predict targeted numerical rating scales (i.e., the semantic responses for harmony predict the targeted rating scale the HILS rather than, e.g., the SWLS) with consistently high correlations. In Study 9 ($N = 477$), the predictions yielded strong correlations to targeted rating scale scores ($r = .58–.72$, $p < .001$; Table 6). Further, results show that the semantic responses of harmony and satisfaction may also be used for predicting rating scores of depression, anxiety, and stress, although with lower correlations compared with the targeted numerical rating scale (Table S9). It is here worth noting that these lower correlations show that the trained scales tend to discriminate between psychological constructs insofar that the semantic responses predict the rating scales of the targeted construct better than other constructs.

Even though the SWLS and the HILS ($r = .80$ to $r = .88$) as well as the GAD-7 and the PHQ-9 ($r = .86$ to $r = .87$) correlate strongly throughout the studies, the semantic responses tend to differentiate among these constructs. In Study 8 and 9 at T2, semantic responses regarding depression are trained to best predict the rating scale of depression, while semantic responses of worry are trained to best predict the rating scale of worry. Similarly, in Studies 3–9, the semantic responses regarding harmony in life are trained to best predict the rating scale of harmony. However, note that this prediction specificity (i.e., that word responses predict the targeted rating scale score better than other rating scale scores) is not consistent with regard to satisfaction. The semantic word responses regarding satisfaction with life are trained to predict rating scales of satisfaction and harmony equally well in Study 5. In Study 6 and 7, where one and three words were required as a response, as well as Study 9 at T2, the satisfaction responses actually predict the rating scale of harmony better than the rating scale of satisfaction. Hence, semantic responses concerning harmony consistently predict the rating scale of harmony better than the rating scale of satisfaction, whereas semantic content of satisfaction appears to predict both rating scales equally well. As the purpose of training is to find the best mathematical fit rather than providing a theoretical explanation, the underlying theoretical reason behind this requires more research. In short, the overall tendency for prediction specificity (e.g., where harmony responses

Table 11
Semantic Questions Responses Differ Significantly Between Related Psychological Constructs

	Harmony versus satisfaction words									Depression versus worry words	
Study: Condition	S3	S4	S5	S6	S7: 1-w	S7: 3-w	S7: 5-w	S9: T1	S9: T2	S8	S9: T2
t-value	2.67	12.71	10.75	7.70	12.36	15.56	12.10	23.66	18.08	16.86	18.43
Cohen's *d*	.28	.73	.63	.56	.66	.83	.76	.81	.85	.85	.84

Note. This table presents semantic *t* tests of the summarized semantic representations of the related psychological constructs. S = study; w = words; T = time; S3 was significant at $p < .01$, all others at $p < .001$; where all except S3, encompass medium to large effect sizes.

Figure 4 (opposite)

predict the HILS better than the SLWS) throughout the studies and across the constructs (except for satisfaction) support the validity of the semantic responses and their semantic representations.

The trained predictions are also robust (Table 7–9). Examining the correlations as a function of participants shows that a sample size of approximately 64 participants is required to reach significant results, although the correlations increase with sample size (see Table 7). Examining the function of generated words reveals that the *first* word accounts for the largest part of the predictive power (see Table 8) and that the ninth and tenth words do not always increase the correlation (see Table 9). To sum up, that semantic responses reliably predict rating scales supports the validity of the semantic measures.

Semantic Similarity Scales Independently Measure Constructs

Semantic similarity scales may be used for measuring subjective states without relying on rating scales. The correlations to rating scales tend to be moderate for unipolar similarity scales and strong for bipolar similarity scales (Table 10, Figure 3). With regard to the satisfaction words, the unipolar satisfaction semantic similarity scale correlates with the SWLS score (in Study 9 T2: $r = .33$), and the bipolar satisfaction semantic similarity scale correlates with the SWLS score to an even higher degree ($r = .60$). Similarly, with regard to the harmony words, the unipolar harmony semantic similarity scale correlates with the HILS score ($r = .43$), and the bipolar harmony semantic similarity scale yields a higher correlation ($r = .65$).

However, it is noteworthy that the semantic similarity scales sometimes correlate higher to a rating scale it is not intended to measure compared with the intended target construct. That is, it could be expected that the semantic similarity scales generated for each construct would exhibit the highest correlation in relation to its specific rating scale. This is the case when it comes to the harmony words and the harmony semantic similarity scale, where the correlation is higher with regard to the HILS ($r = .43$) than with regard to the SWLS ($r = .34$); however, when it comes to the satisfaction words and the satisfaction semantic similarity scale, there is a higher correlation with regard to the HILS ($r = .35$) and not the SWLS ($r = .33$). Similarly, with regard to the worry words and the worry semantic similarity scale, there is a higher correlation when it comes to the GAD-7 ($r = .39$) than to the PHQ-9 ($r = .32$), whereas the depression words and the depression semantic similarity scale do not exhibit a higher correlation when it comes to the PHQ-9 ($r = .29$), but rather when it comes to the GAD-7 ($r = .31$). Hence, there is a lack of clear target specificity among semantic similarity scales and rating scales.

Semantic Similarity Scales Differentiate Between Similar Constructs

We suggest that the less than perfect correlations of semantic similarity scales to rating scales may not necessarily indicate a measurement error in the semantic measures. Instead, semantic similarity scales may measure different qualities that efficiently differentiate between similar constructs, whereas rating scales tend to capture one-dimensional valence. This suggestion was tested by applying an independently trained model of the ANEW (Bradley & Lang, 1999) in order to create semantic predicted valence scales. According to our suggestion, these semantic valence scales correlate significantly stronger with respective rating scales compared with the bipolar similarity scales ($z = 4.32–5.61$, $p < .001$, two-tailed; Lee & Preacher, 2013). Further, using the valence scale as a covariate when training semantic responses to rating scales reduces the correlation considerably (see Table 6). Hence, rating scales are highly influenced by general valence; potentially driven by a general positive or negative attitude to life, rather than distinct answers related to targeted constructs. This interpretation is consistent with findings that the rating scales have strong intercorrelations (see also Kashdan, Biswas-Diener, & King, 2008). On the other hand, the similarity scales based on construct-specific word norms exhibit a lower intercorrelation compared with rating scales. In addition, semantic *t* tests further support the semantic difference between semantic responses by discriminating well between the semantic responses with medium to large effect sizes (see Table 11). This suggests that semantic measures more clearly tap into the targeted construct, which is supported further when describing the constructs using word plots.

Describing Constructs With Keywords

Figures 4–6 provide empirically derived depictions of the studied constructs by plotting words that significantly discriminate in relevant dimensions, including different semantic responses (x-axes) and semantic scales or rating scales (y-axes). The plots tend to conceptualize the constructs in a meaningful way, which is also intuitively consistent with a theoretical understanding of the constructs as outlined in the introduction. The semantic responses concerning satisfaction with life highlight words such as *happy, content*, and *fulfilled*, whereas harmony in life responses highlight *peace*, *calm*, and *balance*. The semantic responses concerning depression and worry are also distinct, where depression is significantly related to words such as *sad*, *down*, and *lonely*, and worry is associated with *anxious*, *nervous*, and *tense*. This supports the construct validity of the semantic measures. Semantic measures empirically describe/define the constructs.

Figure 4. a–f. Semantic measures differentiate between constructs of satisfaction and harmony as shown by plotting significant keywords. The figures also show the independence of constructs using semantic measures (e, f) as compared with numerical rating scales (g, h). On the *x*-axis, words are plotted according to the χ-values from a chi-square test, with a Bonferroni correction for multiple comparisons (Bonf. = Bonferroni line where $\chi = 4.04$, .05 indicates the uncorrected *p* value line, where $\chi = 1.96$). On the *y*-axis, words are plotted according to point biserial correlations ($r = .13$ at the Bonferroni line, and $r = .06$ at the.05 uncorrected *p* value line). More frequent words are plotted with a larger font size with fixed lower and upper limits. The x-axes represent χ-values associated with words generated in relation to satisfaction (blue/left) versus harmony (green/right). The y-axes show significant words related to Semantic Similarity (SS) scales or rating scales; HILS = Harmony in Life Scale; SWLS = Satisfaction with Life Scale; cov = covariate. $N = 477$. See the online article for the color version of this figure.

Figure 5 (opposite)

The ability of measures to offer a good differentiation between constructs may be tested further by examining the number of significant words and their position in the figures when covarying the different scales. Scales that are independent should still yield significant words on the axis where they are being covaried (here the *y*-axis). For example, if the semantic similarity scales of harmony and satisfaction measure the respectively targeted construct with high independence, it should follow that covarying the scales with each other would impact the plotted harmony and satisfaction word responses differently compared with their original positions without a covariate. In contrast, if the scales are *not* measuring the constructs differently/independently, the impact of covarying the scales would have a similar effect on the position of both the harmony and the satisfaction word responses, where both sets of words would be positioned to *not* be significant on the *y*-axis. That is, for semantic similarity scales we hypothesize an independence between the word responses, in which such independence may manifest itself in such a way that word responses for the targeted construct are positioned on the higher end of the *y*-axis, whereas the word responses related to the covaried construct are positioned on the lower end of the *y*-axis. In contrast, when covarying the numerical rating scales, we hypothesize that word responses between constructs will not show independence, and thus not be significant on the *y*-axis anymore.

Indeed, the results show that covarying the harmony semantic similarity scale with the satisfaction semantic similarity scale (or vice versa) reveals a significant independence between these constructs by yielding correlations to relevant words describing the construct; that is, 44 and 37 significant words on the y-axes in Figure 4e and 4f, respectively. This is not the case when covarying the corresponding numerical rating scales; that is, only six and zero significant words on the y-axes in Figure 4g and 4h, respectively. This independence is also found with regard to depression and worry (see Figure 5). When the semantic similarity scales of worry and depression are covaried with each other, there are 35 and 36 significant words on the y-axes in Figure 5e and 5f, respectively. However, when the rating scales of depression and worry are covaried, there are only six and 22 significant words on the y-axes in Figure 5g and 5h, respectively.

Further, the semantic similarity scales actually tap into different aspects than the semantic predicted ANEW valence scales and the rating scales, which becomes clear when described in the word figures. The figures plotting rating scales and semantic predicted valence scales tend to reveal a similar structure, which suggests that rating scales focus on valence. Meanwhile, the semantic similarity scales demonstrate a structure that is more in line with their theoretical conceptualizations. This becomes even more clear when covarying the semantic predicted ANEW valence scales in figures. That is, figures with harmony and satisfaction semantic similarity scales covaried with the semantic predicted ANEW valence scale have 23 and 24 significant words in Figure 6c and 6d, respectively. Meanwhile, the harmony and satisfaction rating scales covaried with the semantic predicted ANEW valence scale reveal zero significant words (Figure 6e–f). This is also the case when it comes to depression and worry: in terms of semantic similarity scales, 25 and 26 words are significant in Figure 6g and 6h, respectively, and in terms of rating scales, zero words are significant (Figure 6i–j). It should be pointed out that this offers further support for the strong, one-dimensional link between rating scales and valence. Overall, these results suggest that similarity scales more clearly differentiate between the psychological constructs compared with rating scales.

Test–Retest

Semantic measures exhibit satisfactory test–retest reliability over 31 days (see Table 12). Trained scales tend to demonstrate moderate correlations for both harmony in life and satisfaction with life. Although unipolar semantic similarity scales demonstrate low test–retest correlations, bipolar semantic similarity scales show moderate correlations for both harmony and satisfaction.

Social Desirability

Overall, the semantic measures, as compared with rating scales, are associated with less social desirability as measured using the Marlowe-Crowne Social Desirability Scale, the short version Form A (Table 13; Reynolds, 1982). The rating scales yield the anticipated positive correlations with the well-being measures and negative correlations with the ill-being measures. Training the semantic content to the social desirability scale only yielded a weak significant relationship for the semantic question of depression. The unipolar semantic similarity scales only yielded a low significant relationship between worry and social desirability. Meanwhile, among the bipolar semantic similarity scales, only harmony–disharmony displayed a weak significant relationship to social desirability. Compared with the rating scales, the semantic predicted ANEW valence scale displayed correlations with similar strength for the ill-being measures and somewhat weaker strength for the well-being measures. Note that all correlations between social desirability and the semantic predicted ANEW valence scale were positive, as positivity in response to the semantic questions tends to relate to social desirability. Further, plotting the words reveals that there are no words significantly related to low or high social desirability scores.

Figure 5. a–h. Semantic measures differentiate between constructs of depression and worry as shown by plotting significant keywords. The figures also show the independence of constructs using semantic measures (e, f) as compared with numerical rating scales (g, h). On the *x*-axis, words are plotted according to the χ-values from a chi-square test, with a Bonferroni correction for multiple comparisons (Bonf. = Bonferroni line where $\chi = 4.04$, .05 indicates the uncorrected *p* value line, where $\chi = 1.96$). On the *y*-axis, words are plotted according to point-biserial correlations ($r = .13$ at the Bonferroni line, and $r = .06$ at the .05 uncorrected *p* value line). More frequent words are plotted with a larger font with fixed lower and upper limits. The x-axes represent χ-values associated with words generated in relation to depression (blue, left) versus worry (red, right). The y-axes show significant words related to Semantic Similarity (SS) scales or rating scales; PHQ-9 = Patient Health Questionnaire-9; GAD-7 = Generalized Anxiety Disorder scale-7 (GAD-7); cov. = covariate; $N = 477$. See the online article for the color version of this figure.

Figure 6 (opposite)

Table 12
Semantic Questions Demonstrate Satisfactory Test–Retest Reliability for the Well-Being Measures in Study 9 Between T1 and T2

Semantic questions	Pearson's *r*
Unipolar semantic similarity scales	
Hw: Unipolar at T1 and T2	.24***
Sw: Unipolar at T1 and T2	.20***
Bipolar semantic similarity scales	
Hw: Bipolar at T1 and T2	.52***
Sw: Bipolar at T1 and T2	.48***
Semantic predicted valence scales	
Hw: Valence at T1 and T2	.55***
Sw: Valence at T1 and T2	.52***
Trained models	
T1 Hw: HILS and T2 Hw: HILS	.49***
T1 Sw: SWLS and T2 Sw: SWLS	.45***
Rating scales	
HILS at T1 and T2	.71***
SWLS at T1 and T2	.82***

Note. $N = 477$. Hw = Harmony words; Sw = Satisfaction words; H-SSS = Harmony-Semantic Similarity Scale; S-SSS = Satisfaction-Semantic Similarity Scales; Hw: Bipolar = Harmony minus Disharmony Semantic Similarity Scale; Sw: Bipolar = Satisfaction minus Dissatisfaction-Semantic Similarity Scale; HILS = Harmony in Life Scale; SWLS = Satisfaction with Life Scale.
*** $p < .001$.

Discussion on Reports Regarding Subjective States

The results reveal that semantic measures may be used for measuring, differentiating, and describing subjective experiences of psychological constructs, including both well-being and ill-being. It is demonstrated that trained semantic responses predict rating scales scores with a high level of accuracy; which hold throughout seven studies involving differences in terms of required word responses (with 1, 3, 5, or 10 descriptive words, or free text as presented in the OSM) and varying levels of detailed instructions. In addition, using semantic similarity scales enables measuring, differentiating and describing psychological constructs independent from rating scales. It is also shown that semantic measures show satisfactory test–retest reliability. Further, the four semantic questions, as compared with the corresponding rating scales, appear less susceptible to social desirability. Arguably, this might be because they promote a more personal and thus honest account of one's state of mind.

Semantic Measures Versus Rating Scales

In our studies, semantic similarity scales appears to exhibit higher construct specificity and independence to similar constructs compared with rating scales. The word figures exhibit target specificity and independence when it comes to semantic similarity scales but not rating scales, even though there is a lack of target specificity among correlations between semantic similarity scales and respective rating scales. In addition, this target specificity and independence were found for both descriptive words in relation to harmony, satisfaction, depression, and worry, as well as text responses in relation to harmony and satisfaction (see OSM). Overall, the results support that semantic measures demonstrate high construct validity where they clearly tap into the to-be-measured construct, whereas rating scales to a greater extent appear to be tapping into a common construct relating to valence.

Valence-Focused Rating Scales and Construct-Specific Semantic Measures

What rating scales might have in common is that they more strongly tap into valence, whereas semantic similarity scales are better at differentiating between constructs. The hypothesis that rating scales are highly valence-driven is supported empirically by the results showing that covarying the semantic predicted ANEW valence scales when training the semantic responses to rating scales largely reduces the correlations. Similarly, our results show that the semantic predicted ANEW valence scales of each semantic response have a higher correlation with the rating scales as compared with respective semantic similarity scale. Hence, the affective valence of the generated words is strongly related to the rating scales.

In addition, the word figures demonstrate that semantic similarity scales, as compared with rating scales, are more distinct from the semantic predicted ANEW valence scales. That rating scales exhibit none or few significant words when being covaried with the semantic predicted ANEW valence scales, as well as with each other, is also in accordance with previous findings, including the high intercorrelation among rating scales and the lack of target specificity between correlations of rating scales and semantic similarity scales.

From a theoretical perspective, we argue that rating scales and valence have one-dimensionality in common. That is, valence is conceptualized as ranging from pleasant to unpleasant

Figure 6. a–j. Compared with rating scales, semantic measures do a better job differentiating between mental health constructs beyond valence. On the *x*-axis, words are plotted according to the χ-values from a chi-square test, with a Bonferroni correction for multiple comparisons (Bonf. = Bonferroni line where $\chi = 4.04$, .05 indicates the uncorrected *p* value line, where $\chi = 1.96$). On the *y*-axis, words are plotted according to point-biserial correlations ($r = .13$ at the Bonferroni line, and $r = .06$ at the .05 uncorrected *p* value line). More frequent words are plotted with a larger font with fixed lower and upper limits. The x-axes represent χ-values associated with words generated in relation to a, c-f) satisfaction (blue, left) versus harmony (green, right) and b, g-j) depression (blue, left) versus worry (red, right). The y-axes in Figures c-j are covaried (cov.) with the semantic predicted (SP) ANEW valence scale; SS = Semantic Similarity; HILS = Harmony in Life Scale; SWLS = Satisfaction with Life Scale; PHQ-9 = Patient Health Questionnaire-9; GAD-7 = Generalized Anxiety Disorder scale-7. $N = 477$. See the online article for the color version of this figure.

Table 13
Semantic Questions Show Lower Social Desirability Than Rating Scales

Correlated constructs	Training to MC-SDS-FA	Valence SP scales	Unipolar SS scales	Bipolar SS scales	Rating scales
MC-SDS-FA					
Harmony	ns	.11*	ns	.11*	.19***
Satisfaction	ns	.11*	ns	ns	.19***
Worry	ns	.15***	−.10*	*NA*	−.16***
Depression	.08*	.13**	ns	*NA*	−.14**

Note. Study 9 at T2, $N = 477$. SP = Semantic Predicted; SS = Semantic Similarity; ns = not significant; MC-SDS-FA = The Marlowe-Crowne Social Desirability Scale the short version Form A.
* $p < .05$. ** $p < .01$. *** $p < .001$.

(i.e., in the construction of the ANEW, words were rated on a picture-based scale depicting pictures of happy, smiling to unhappy, frowning characters; Bradley & Lang, 1999) This is arguably similar to the one-dimensional response format often used for rating scales. Response formats for rating scales tend to focus on the respondent's positive or negative stance on various statements (although the scale may differ, such as ranging from *strongly disagree* to *strongly agree* or *not at all* to *nearly every day*). This may potentially explain the lack of clear specificity among rating scales and semantic similarity scales.

Overall, the results suggest that rating scales are more focused on valence compared with semantic similarity scales, whereas semantic similarity scales are more focused on the to-be-measured construct compared with rating scales. Hence, the results indicate that semantic similarity scales, independent from rating scales, may be used for measuring, differentiating and describing participants' subjective experience of harmony in life, satisfaction with life, depression and worry with high validity and reliability.

Limitations, Future Research, and Overall Conclusions

Answering semantic questions seems to require more time and effort for participants compared with rating scales. In the studies on reports regarding external stimuli, the semantic questions conditions, as compared with the rating scales conditions, took longer to complete for participants (see Table 2). In terms of effort, the semantic questions conditions yield lower percentages of participants completing the study, presumable due to the effort required for generating semantic answers. Even though the randomization procedure should divide 50% of the participants in the semantic questions condition, there were only 45% in Study 1 and 42% in Study 2 completing this condition. However, it should be noted that semantic questions conditions may be rendered less time-consuming by requesting fewer words; for example, we found that it is possible to achieve good results by only requiring *one* word, as well as using short instructions in reports regarding subjective states. In addition, the relatively higher level of effort needed for semantic questions might be one of the strengths of this approach. That is, semantic questions may lead to more thought through answers. In contrast, instructions for rating scales often ask respondents to *not* think for too long on each question, but instead answer with what first comes to mind.

The current studies focus on psychological constructs relating to well-being and mental health problems; however, future studies could study a broader range of psychological constructs and contexts. Further, the current semantic questions allow the respondent to decide for him- or herself what is important for the to-be-measured construct. However, in terms of mental health diagnoses, future research could examine to what extent these semantic questions cover important diagnostic criteria outlined in manuals such as the *DSM–5*.

Moreover, the studies on subjective states only include self-report rating scales, whereas future studies should evaluate these using objective measures. In addition, whereas the current studies compare semantic measures with rating scale methodologies, future studies may also compare them with interview techniques (e.g., the International Neuropsychiatric Interview; MINI; Sheehan et al., 1998). Finally, future studies could also explore potential advantages with using semantic representations based on other algorithms, such as COALS, as well as word frequency strategies, such as LIWC.

To sum up, our results from experiments based on both external stimuli and subjective states show that semantic measures are competitive, or better, compared with numerical rating scales when evaluated both in terms of validity and reliability. Semantic measures address limitations associated with rating scales, for example by allowing unrestricted open-ended responses. Importantly, semantic similarity scales enable measuring constructs independent of rating scales, and we demonstrated that they are good at differentiating even between similar constructs. Trained semantic scales may also be used for closely studying the relationship between texts/words and numeric values. This is particularly valuable when numeric values represent objective outcomes rather than subjective rating scales, when studying a particular characteristic of a word response, such as valence, or in situations when it is difficult to construct a representative word norm (such as in the case of social desirability). Future research should investigate to what extent these results generalize to other psychological constructs, situations and contexts.

That semantic measures are able to *measure, differentiate*, and *describe* psychological constructs in two different paradigms demonstrate their potential to improve the way we quantify individual states of mind, and thus our understanding of the human mind. Therefore, we argue that semantic measures offer an alternative method for quantifying mental states in research

and professional contexts (e.g., surveys, opinion polls, job recruitment, etc.).

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: American Psychiatric Association.

Arvidsson, D., Sikström, S., & Werbart, A. (2011). Changes in self and object representations following psychotherapy measured by a theory-free, computational, semantic space method. *Psychotherapy Research, 21,* 430–446. http://dx.doi.org/10.1080/10503307.2011.577824

Axelson, D. A., & Birmaher, B. (2001). Relation between anxiety and depressive disorders in childhood and adolescence. *Depression and Anxiety, 14,* 67–78. http://dx.doi.org/10.1002/da.1048

Bartlett, M. S., Viola, P. A., Sejnowski, T. J., Golomb, B. A., Larsen, J., Hager, J. C., & Ekman, P. (1996). Classifying facial action. *Advances in Neural Information Processing Systems, 36,* 823–829.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Brown, T. A. (2007). Temporal course and structural relationships among dimensions of temperament and *DSM–IV* anxiety and mood disorder constructs. *Journal of Abnormal Psychology, 116,* 313–328.

Brown, T. A., Chorpita, B. F., & Barlow, D. H. (1998). Structural relationships among dimensions of the *DSM–IV* anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. *Journal of Abnormal Psychology, 107,* 179–192. http://dx.doi.org/10.1037/0021-843X.107.2.179

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6,* 3–5. http://dx.doi.org/10.1177/1745691610393980

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24,* 349–354. http://dx.doi.org/10.1037/h0047358

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49,* 71–75. http://dx.doi.org/10.1207/s15327752jpa4901_13

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105,* 399–412. http://dx.doi.org/10.1111/bjop.12046

Ekman, P. (1992). Are there basic emotions? *Psychological Review, 99,* 550–553. http://dx.doi.org/10.1037/0033-295X.99.3.550

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes, 25,* 285–307. http://dx.doi.org/10.1080/01638539809545029

Garcia, D., & Sikström, S. (2013). Quantifying the semantic representations of adolescents' memories of positive and negative life events. *Journal of Happiness Studies, 14,* 1309–1323. http://dx.doi.org/10.1007/s10902-012-9385-8

Golub, G., & Kahan, W. (1965). Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics, Series B*. Numerical Analysis, *2,* 205–224.

Gustafsson Sendén, M., Lindholm, T., & Sikström, S. (2014). Biases in news media as reflected by personal pronouns in evaluative contexts. *Social Psychology, 45,* 103–111.

Gustafsson Sendén, M., Sikström, S., & Lindholm, T. (2015). 'She' and 'He' in news media messages: Pronoun use reflects gender biases in semantic contexts. *Sex Roles, 72,* 40–49. http://dx.doi.org/10.1007/s11199-014-0437-x

Karlsson, K., Sikström, S., & Willander, J. (2013). The semantic representation of event information depends on the cue modality: An instance of meaning-based retrieval. *PLoS ONE, 8,* e73378. http://dx.doi.org/10.1371/journal.pone.0073378

Kashdan, T. B., Biswas-Diener, R., & King, L. A. (2008). Reconsidering happiness: The costs of distinguishing between hedonics and eudaimonia. *The Journal of Positive Psychology, 3,* 219–233. http://dx.doi.org/10.1080/17439760802303044

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological Methods, 21,* 507–525. http://dx.doi.org/10.1037/met0000091

Kessler, R. C., Chiu, W. T., Demler, O., Merikangas, K. R., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month *DSM–IV* disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry, 62,* 617–627. http://dx.doi.org/10.1001/archpsyc.62.6.617

Kjell, O. N. E. (2011). Sustainable well-being: A potential synergy between sustainability and well-being research. *Review of General Psychology, 15,* 255–266. http://dx.doi.org/10.1037/a0024603

Kjell, O. N. E., Daukantaitė, D., Hefferon, K., & Sikström, S. (2015). The harmony in life scale complements the satisfaction with life scale: Expanding the conceptualization of the cognitive component of subjective well-being. *Social Indicators Research*. Advance online publication.

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals, 32,* 1–7. http://dx.doi.org/10.3928/0048-5713-20020901-06

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine, 16,* 606–613. http://dx.doi.org/10.1046/j.1525-1497.2001.016009606.x

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50,* 537–567. http://dx.doi.org/10.1146/annurev.psych.50.1.537

Kwantes, P. J., Derbentseva, N., Lam, Q., Vartanian, O., & Marmurek, H. H. C. (2016). Assessing the Big Five personality traits with latent semantic analysis. *Personality and Individual Differences, 102,* 229–233. http://dx.doi.org/10.1016/j.paid.2016.07.010

Landauer, T. K. (1999). Latent semantic analysis: A theory of the psychology of language and mind. *Discourse Processes, 27,* 303–310. http://dx.doi.org/10.1080/01638539909545065

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240. http://dx.doi.org/10.1037/0033-295X.104.2.211

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes, 25,* 259–284. http://dx.doi.org/10.1080/01638539809545028

Landauer, T. K., McNamara, D., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis. University of Colorado Institute of cognitive science series*. Hillsdale, NJ: Erlbaum.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition and Emotion, 24,* 1377–1388. http://dx.doi.org/10.1080/02699930903485076

Lee, I., & Preacher, K. (2013). *Calculation for the test of the difference between two dependent correlations with one variable in common* [Computer software]. Retrieved from http://quantpsy.org

Li, C. (2008). The philosophy of harmony in classical Confucianism. *Philosophy Compass, 3,* 423–435. http://dx.doi.org/10.1111/j.1747-9991.2008.00141.x

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

Löwe, B., Decker, O., Müller, S., Brähler, E., Schellberg, D., Herzog, W., & Herzberg, P. Y. (2008). Validation and standardization of the generalized anxiety disorder screener (GAD-7) in the general population.

Medical Care, 46, 266–274. http://dx.doi.org/10.1097/MLR.0b013e318160d093

Martin, A., Rief, W., Klaiberg, A., & Braehler, E. (2006). Validity of the brief patient health questionnaire mood scale (PHQ-9) in the general population. *General hospital psychiatry, 28,* 71–77.

McNamara, D. S. (2011). Computational methods to extract meaning from text and advance theories of human cognition. *Topics in Cognitive Science, 3,* 3–17. http://dx.doi.org/10.1111/j.1756-8765.2010.01117.x

Muntingh, A. D. T., van der Feltz-Cornelis, C. M., van Marwijk, H. W. J., Spinhoven, P., Penninx, B. W. J. H., & van Balkom, A. J. L. M. (2011). Is the Beck Anxiety Inventory a good tool to assess the severity of anxiety? A primary care study in the Netherlands Study of Depression and Anxiety (NESDA). *BMC Family Practice, 12,* 66. http://dx.doi.org/10.1186/1471-2296-12-66

Nakov, P., Popova, A., & Mateev, P. (2001). Weight functions impact on LSA performance. *EuroConference RANLP,* 187–193.

Neuman, Y., & Cohen, Y. (2014). A vectorial semantics approach to personality assessment. *Scientific Reports, 4,* 4761. http://dx.doi.org/10.1038/srep04761

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349,* aac4716. http://dx.doi.org/10.1126/science.aac4716

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45,* 867–872. http://dx.doi.org/10.1016/j.jesp.2009.03.009

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., . . . Seligman, M. E. P. (2014). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*. Advance online publication.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count: LIWC 2001*. Mahwah, NJ: Erlbaum.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language. use: Our words, our selves. *Annual Review of Psychology, 54,* 547–577. http://dx.doi.org/10.1146/annurev.psych.54.101601.145041

Reynolds, W. M. (1982). Development of reliable and valid short forms of the Marlowe-Crowne social desirability scale. *Journal of Clinical Psychology, 38,* 119–125. http://dx.doi.org/10.1002/1097-4679(198201)38:1<119::AID-JCLP2270380118>3.0.CO;2-I

Rohde, D. L. T., Gonnerman, L. M., & Plaut, D. C. (2005). *An improved model of semantic similarity based on lexical co-occurrence*. Unpublished manuscript. Retrieved from http://tedlab.mit.edu/~dr/Papers/RohdeGonnermanPlaut-COALS.pdf

Roll, M., Mårtensson, F., Sikström, S., Apt, P., Arnling-Bååth, R., & Horne, M. (2012). Atypical associations to abstract words in Broca's aphasia. *Cortex, 48,* 1068–1072. http://dx.doi.org/10.1016/j.cortex.2011.11.009

Sarwar, F., Sikström, S., Allwood, C. M., & Innes-Ker, Å. (2015). Predicting correctness of eyewitness statements using the semantic evaluation method (SEM). *Quality & Quantity: International Journal of Methodology, 49,* 1735–1745. http://dx.doi.org/10.1007/s11135-014-9997-7

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., . . . Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M. I. N. I.): The development and validation of a structured diagnostic psychiatric interview for *DSM–IV* and ICD-10. *The Journal of Clinical Psychiatry, 59,* 22–33.

Sinclair, S. J., Siefert, C. J., Slavin-Mulford, J. M., Stein, M. B., Renna, M., & Blais, M. A. (2012). Psychometric evaluation and normative data for the depression, anxiety, and stress scales-21 (DASS-21) in a nonclinical sample of U.S. adults. *Evaluation & the Health Professions, 35,* 259–279. http://dx.doi.org/10.1177/0163278711424282

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine, 166,* 1092–1097. http://dx.doi.org/10.1001/archinte.166.10.1092

Watson, D., Clark, L. A., & Carey, G. (1988). Positive and negative affectivity and their relation to anxiety and depressive disorders. *Journal of Abnormal Psychology, 97,* 346–353. http://dx.doi.org/10.1037/0021-843X.97.3.346

Wittchen, H. U., Kessler, R. C., Beesdo, K., Krause, P., Höfler, M., & Hoyer, J. (2002). Generalized anxiety and depression in primary care: Prevalence, recognition, and management. *The Journal of Clinical Psychiatry, 63,* 24–34.

Received January 25, 2017
Revision received March 28, 2018
Accepted April 10, 2018 ■