Department of Psychology

Bachelor Thesis

Investigating Validity of Semantic Measures vs Rating Scales in Assessing Personality

Sofie Gustafsson & Henrietta Plate

Bachelor Thesis Spring 23

Supervisor: Sverker Sikström

Examiner: Magnus Lindén

Abstract

Within personality research, self-report questionnaires are a common approach. This study takes aim at investigating whether self-report questionnaires are enough, or if semantic measures, through Natural Language Processing, could be a substitute or complementary method, in assessing personality. Based on the Five Factor Model of personality, this study has been divided into two phases. Participants originating from the U.S. were instructed to either describe and rate their own or someone else's personality (Phase 1, *N*=264), or read personality narratives from Phase 1 (Phase 2, *N*=399) then, in both phases, participants were asked to answer a semantic question as well as a IPIP-NEO rating scale. Prediction scores from the two phases were used to analyze semantic measures in comparison to rating scales. The results suggest that semantic measures, on their own, categorize personality traits more accurately (53%) than rating scales (44%). To conclude, complementary approaches while assessing personality have shown to be of great value. Future research would be benefitted by investigating the possibility of applying this method in other fields of psychology, in favor of further assessing the method's validity, and determining whether it can offer novel understandings of constructs.

Keywords*: Rating scales, semantic measures, Five Factor Model of personality, The Big Five, IPIP-NEO 30, Natural Language Processing*

Validity of Semantic Measures

The following study explores whether semantic measures with open ended questions, analyzed using Natural Language Processing (NLP), could be a complementary method to self-report questionnaires with close ended questions, or maybe even a substitute, for assessing personality traits.

Five Factor Model of personality

Personality tests and research are widely used in a myriad of fields, for example within clinical psychology and recruitment. Personality constructs are commonly measured using self-report questionnaires and most focus on personality traits as expressed by the Five Factor Model of personality (FFM), also known as the "Big Five" personality traits, a widely researched and accepted model of personality. The model emerged from a series of factor-analytic studies in the 1950s and 1960s, and has since been validated by numerous studies across different cultures and in various languages (McCrae, 1989, McCrae & Costa, 1997). The five traits of personality in the FFM are the following: *Openness*: characterized by imagination, creativity, and a willingness to try new things, *Conscientiousness*: characterized by organization, responsibility, and reliability, *Extraversion*: characterized by sociability, talkativeness, and assertiveness, *Agreeableness*: characterized by kindness, compassion, and a tendency to get along with others, *Neuroticism*: characterized by negative emotions, anxiety, and moodiness (Costa & McCrae, 1992).

The Big Five became one of the dominating trait theories after empirical studies demonstrated that the traits being assessed by psychological questionnaires were very much in conjunction with the lexical Big Five Factors. It became evident that in creating these questionnaires personality psychologists only defined, cataloged and formalized what had already been implicit in layman conception of personality (McCrae, 1989).

The extensive acceptance of the Big Five ushered more research in the field which contributed to valuable advances in personality trait psychology. The model has proven to be an efficient tool for characterizing individual differences across dimensions with high reliability as well as validity (Digman, 1990).

Natural Language and Personality Assessment.

Assessing personality is inherently tied with natural language. Every culture and language has specific words that illustrate and describe individual differences in personality, and an important part of socializing with other humans is learning these specific terms and being able to accurately apply them to oneself and others. Traits are abstractions that cannot be measured directly, instead they must be inferred from complex patterns of behavior expressed using natural language, typically by responding to questionnaires (McCrae & Costa, 1997). Until recently we have been relying on humans to make these inferences, even in terms of technical judgments such as diagnosing psychiatric conditions. However, following technological advances, computers are now able to accurately characterize abstractions expressed through natural language, using Natural Language Processing (NLP) and machine learning (ML).

Recent research in the field has found that NLP AIs indeed perform better than humans in analyzing natural language and performing correct categorizations of abstractions, such as emotions, further validating the method (Kjell et al., 2019).

Taking this into consideration, the aim of our research is to determine whether semantic measures using NLP is applicable in the field of personality trait psychology to further validate the method and demonstrate the promise of this method as an alternative to rating scales in many different fields of psychology.

The relationship between personality and language use has been highlighted by many researchers and systematic associations between the two have been identified. This has been done in a variety of contexts such as directed writing assignments (Kwantes et al., 2016), large-scale analysis of language use among bloggers (Yarkoni, 2010), large-scale text analysis from three vastly different sample groups (Pennebaker & King, 1999), and naturalistic recordings of everyday speech (Mehl, Gosling, & Pennebaker, 2006). The results of these studies have consistently verified theoretical predictions and identified large numbers of associations between linguistic style and personality, suggesting that using semantic measures would be a meaningful way of exploring personality constructs.

Personality influences the way people use language as well as what they choose to write or talk about (Yarkoni, 2010). Therefore open-ended questions about personality analyzed using NLP could be a viable alternative to the widespread use of close-ended rating scales when it comes to evaluating, predicting and describing the Big Five and other personality constructs. This could allow for a more comprehensive understanding of individuals and their unique minds as well as give way for new explorations of personality and the novel ways in which it may be described. It could also be used as a way of partially avoiding the perpetual social desirability and acquiescence biases that are closely linked to rating scales (Kjell et al., 2019).

Research has shown that responses vary greatly depending on the setting/context in which a questionnaire is answered. Close ended rating scales are frequently used in clinical settings such as psychiatric evaluations, as well as in professional settings, such as job recruitment. Both of these fields are subject to response biases in their evaluations of personality since the high pressure environment is likely to make the person of interest choose answering options that shed them in the best light (Hinz et al., 2007).

Rating scales vs. semantic measurements

Rating scales such as IPIP-NEO are frequently used in psychology research and are well-established measures of personality in the context of the Big Five. They are quick and easy to administer making them useful in time-sensitive research contexts and studies with large sample sizes. Since they provide clear and straightforward measures of personality they have high face validity as well as high perceived credibility (Uher, 2018; Costa & McCrae, 1992)

However, the method is known for having a high potential for social desirability bias, meaning participants may respond in ways that they perceive is socially desirable rather than accurately reflecting upon their true selves. Rating scales also provide very limited views of personality as they generate a general overview of traits rather than a detailed, insightful understanding of personality or novel aspects of the constructs (Uher, 2018).

Computational language assessment offers fine-grained analysis of constructs, NLP AIs can analyze large quantities of text and extract detailed information about psychological constructs based on patterns and word usage. It offers an unobtrusive measurement, as participants provide written text responses, meaning the potential for social desirability bias is significantly less than with rating scales. Since NLP AIs are capable of large-scale analysis, it can handle large amounts of text data, it is a suitable method for studying large populations, in the same way that rating scales are (Kjell et al., 2019).

Computational language assessment may however miss aspects of personality that are not easily captured in text, such as body language or verbal cues. The results generated by NLP AI models can be influenced by the data and assumptions used to train the model, and can also be affected by a limited scope. Semantic measures are often based on a limited set of words and concepts, and may not capture the full range of personality traits and individual differences (Pilehvar & Camacho-Collados, 2020; McDaniel et al., 2016).

There are still debates ongoing concerning the validity of computational language assessment as a method for measuring psychological constructs, which is what this study aims to investigate. Recent research has led to further development of this computational method aiming to accurately both measure and describe constructs while also capturing how individuals naturally answer questions about subjective states (Kjell et al., 2019).

Present study

Current psychology research is investigating and impugning to what extent rating scales can be relied upon in determining and assessing psychological constructs. The present study reposes on the work by Kjell and colleagues (2019), in which the validity of semantic measures in comparison to that of rating scales is evaluated with regards to interpreting a person's true state of mind, in terms of subjective experiences such as emotions. While the research shows promising results regarding the application of semantic measures instead of or in conjunction with rating scales, the applicability of the method in different fields of psychology still needs to be investigated, such as the field of personality research. Despite growing recognition, semantic measures are still relatively new and have not been as extensively validated as traditional measures of personality, such as self-report questionnaires, which is why more research needs to be conducted.

The present study assesses the predictive properties of semantic measures with regards to the Big Five personality traits and explores whether open-ended text responses analyzed using Natural Language Processing (NLP) can lead to more accurate, all encompassing descriptions of the personality constructs as well as predictions of the five traits compared to those generated with rating scales. Using NLP the text based descriptions of personality are contrasted with their reflection in a numerical scale evaluating the Big Five traits, namely IPIP-NEO personality questionnaire. The following hypotheses are explored:

1. When assessing the Big Five personality constructs, semantic measures analyzed using NLP have higher predictive properties than the IPIP-NEO rating scale.
2. When assessing the Big Five personality constructs, semantic measures and IPIP-NEO combined have higher predictive accuracy than either method isolated.

Methods

Participants

In total, 290 participants completed the questionnaire for Phase 1 and 264 of these answers were included in the analysis. These participants consisted of 141 females, 116 males, 6 non-binary/third gender and 1 preferred not to say. The age ranges from 19 to 82 years (M = 39.22, SD = 44.54).

In Phase 2, a total of 490 participants completed the questionnaire and 399 were included in the analysis. These participants consisted of 234 females, 149 males, 7 non-binary/third gender and 9 preferred not to say. The ages range from 18 to 82 years (*M* = 40.81, *SD* =13.97).

Criterias for participation in the study consisted of having English as your first language and being 18 years of age or older.

Materials

To gather information about personality quantitatively, a scale measurement was required. The Five Factor Model (FFM) is a form of conceptualization of personality subsisting of five trait domains - agreeableness, conscientiousness, extraversion, neuroticism and openness to experience (Kajonius & Johnson, 2019). A rating scale instrument called IPIP-NEO 30 was used to conduct the data collection and consists of 30 scale items such as “Get stressed out easily” (neuroticism) and “Avoid crowds” (reversed extraversion). The questions are divided into the five factor model traits domains where each domain includes six questions each. Some of the questions are reversed for the sake of the traits to generally be scored high or low, which will facilitate the data analysis (ibid. 2019). Participants were given a five point rating scale between *1 = never* to *5 = always* to answer within.

Latent Semantic Analysis (LSA) is an approach within Natural Language Processing which was also included in the study. The method of LSA is the expectation of a specific word to belong within a context that could bring deeper information about the meaning of the word to surface (Kjell et al., 2019).

Procedure

The study was divided into two phases. In Phase 1 the data was collected using a questionnaire made in Qualtrics. The participants were recruited through Prolific and received compensation for their time spent completing the questionnaire, calculated for £7.5/hour. The questionnaire consisted of two parts - a qualitative and a quantitative, split into two conditions. The parts had the same structure for both conditions. Each participant needed to complete both parts, but only one of the conditions.

The conditions for the study was for participants to either describe their own personality in the “self condition”, or describe the personality of a person they know that matches one of the Big Five traits - neuroticism, extraversion, openness, agreeableness or conscientiousness in the “other condition”. Even though the participants were assigned a trait, they were asked to describe the person's personality as a whole, and not base their descriptions solely on the given trait. Important to note is that the participants were not given the name of the trait itself, but rather a description of it.

In part one, the participants were asked to, qualitatively, with words, describe their own or another person's personality using 300 characters or more. The participants were also asked to write five keywords that best captured their own or the others personality. For the “other” condition, the participants were asked not to use words from the trait descriptions they were given, for example “neurotic” or “extroverted”. In part two the participants were asked to answer 30 questions on a rating scale concerning their own or the others personality. The question items are based on IPIP-NEO. Apart from the 30 IPIP-NEO based questions, the questionnaire also included three control questions to ensure that the participants read the questions thoroughly.

The distribution of the questionnaires was 50% chance of getting the “self” condition, and 50% of getting the “other” condition. Within the “other” condition, the chances of getting one of the five traits were equal. The questionnaire took approximately 10 minutes to complete.

In Phase 2, the participants were asked to read text answers collected from Phase 1. Thereafter sum the personality they read about in five keywords and answer the same 30 IPIP-NEO questions in a rating scale, same as in phase 1. The aim of this questionnaire was to examine any possible discrepancy between the personality reading for the first and second phase.

The questionnaire in Phase 2 was equally distributed between the “self” and “other” condition as in Phase 1 and took approximately 5 minutes to complete. The participants in the Phase 2 were also compensated with £7.5/hour, same as in Phase 1.

In both phases participants were asked to answer basic demographic questions including their age, gender and educational level, as well as questions about life satisfaction and how well they believe in their skills of describing themselves. The participants were also asked to complete a language comprehension test. The reason for this was to establish their understanding of the questionnaire.

Study Design

The study is conducted cross-sectionally, meaning that the data collected will be analyzed comparing the results of Phase 1 and 2. The validity of the study design should be considered in relation to that both the quantitative and qualitative parts can be compared to identify the participants' understanding of the assignment and contribution of information about their own or others personalities. The study's reliability would be considered in relation to the fact that the method has been used before in other studies and provided strong results at different points in time and within different fields of psychology research.

Considering the ethical aspects no sensitive information was collected or distributed. If the participants would experience discomfort while describing their own or others personality they had the opportunity to cancel their participation at any time. Participants could not access the rest of the questions if they did not actively consent in the beginning of the questionnaires.

Data Analysis

Pre-processing of the Data

Phase 1 consisted of a total amount of 264 participants and Phase 2 of a total amount of 399 participants. Answers were excluded from the analysis of both phases if the participant answered the control questions incorrectly, answered the free text or keywords frivolously or if they answered the scale questions with the same scale number for every item. Issues with the keywords were that some participants answered with sentences, some did not fully understand what the study was reaching for considering personality descriptive terms and some didn't answer seriously, with examples like “dolphin” or “baked goods”.

Texts from Phase 1 were either disqualified for usage in Phase 2 or used in their original form. The only alterations made, in a few of the usable texts, were that serious misspellings were corrected if the text was qualitatively valuable. A text could for example be corrected as follows.

Original:

“the Person ive chosen is a man by the name of Greg he is, a very focused, and prepared man, he was the treausre of tla church i used to attend, and also in charge of constructing misson trips. Hes very compassionate, and very kind. He knows how to navigste himself as well as a team of people. while on mission trips he is very open to the ideas of others, flexible to their individual needs, and tending to the atmosphere of the team as a whole.”

Corrected:

"The person I’ve chosen is a man by the name of Greg. He is a very focused, and prepared man, he was the treasure of the church I used to attend, and also in charge of constructing mission trips. He’s very compassionate, and very kind. He knows how to navigate himself as well as a team of people. While on mission trips he is very open

to the ideas of others, flexible to their individual needs, and tending to the atmosphere of the team as a whole."

These kinds of corrections were made with the intention of making the texts more understandable for the readers in Phase 2. Though it has been shown to be counterproductive, due to SemanticExcel's inability to match the texts from Phase 1 and Phase 2 if they are not completely identical.

Statistical Analyses

The words generated from the participants' semantic responses were quantified following the methods described in Kjell et al (2019), creating semantic representations. The representations are contained in vectors where words are assigned numerical values within a semantic space. The semantic space is a matrix that represents how all words relate to one another across dimensions. A general semantic space, generated from a larger data-set with a basis in inter-text word co-occurrences has been applied in the text analysis as a model before adding on the semantic data collected in this study.

The predicted categorization of the Big Five personality traits was generated using multinomial logistics regression, conducted in SemanticsExcel.com (Sikström et al., 2020) using code written in MATLAB. The categorization was based either on the semantic dimensions generated from the keywords, or on the IPIP-NEO rating scales or both combined (see Table 1). Multiple linear regression analysis was used to predict rating scales on the basis of the semantic representations. The predictive validity of the semantic responses was examined using semantic numeric correlations, where the semantic method was tested in relation to rating scales, see Table 2 (Kjell et al., 2019). For example: in the "self" condition, the semantic responses by participants that have been primed with a certain trait ("other") is used to train a predictive model in guessing how the non-primed participants are going to apply language to describe themselves.

Word clouds were created using Semantic Excel (see figures 1-5). To run semantic t-tests, summarization of the keywords on semantic representation from one of the traits is normalized as a vector to the length of one, which then is used as a semantic representation for word responses for the rest of the four personality traits (Kjell et al., 2020). The clouds represent the words with the highest t-values, which also were statistically significant.

Results

Correct Categorizations of Big Five Traits - As seen in Table 1

For both phases combined (All), the percentage of correct categorizations of personality traits was 44% based on rating scales (RS), 53% based on semantic responses (Words) and 54% based on both methods combined (Words + RS).

In Phase 1 the percentage of correct categorizations was 50% for RS, 58% for Words and 62% for Words + RS.

In Phase 2, the percentage of correct categorizations was 40% for RS, 49% for Words and 48% for RS+Words.

To summarize, rating scales and words combined (RS + Words) was overall the most successful in categorizing the personality traits, however only marginally better than semantic measures (Words) as a method on its own. Rating scales have the smallest percentage of correct categorizations across all phases.

Table 1.

Ratio of correct classifications (N(correct)/N(total)) based on semantic response (words) and rating scales (RS) or both combined (RS+ Words), divided into phases. The "Correct" column should be read as percentages.

Scale	Phase	*N*	*Correct*
RS	All	298	0.44
Words	All	298	0.53
RS+Words	All	298	0.54
RS	1	125	0.50
Words	1	125	0.58
RS+Words	1	125	0.62
RS	2	173	0.40
Words	2	173	0.49
RS+Words	2	173	0.48

Correlations Between Semantic Responses and Rating Scale Scores in Predicting Traits

In the correlation matrix (see table 2) the predictive validity of the methods has been tested. The semantic measure model (W1/W2) was in both conditions combined significantly more accurate in predicting personality traits compared to the rating scale (RS1/RS2) for every trait except neuroticism (p = 0.0932) and extraversion (p = 0.2609). The correlations were consistently high ranging from ($0.43 < r < 0.55$) in the semantic measure model, and ranging from ($0.15 < r < 0.49$) in the rating scale model.

In the "self" condition, where participants were instructed to describe themselves, the correlations range from ($0.12 < r < 0.51$) in the rating scale model and ($0.23 < r < 0.47$) in the semantic measure model. In this condition the differences between the two groups were not significant for any of the five traits.

In the "other" condition, where participants were instructed to describe someone else, the semantic measure model had significantly higher correlations for all of the traits. The correlations ranged from ($0.50 < r < 0.65$) in the semantic measure model and ($0.03 < r < 0.45$) in the rating scale model.

Table 2.

A correlation matrix showing Pearson correlation coefficient (r) for predictions of personality traits between semantic responses (W) and Rating Scales (RS). It illustrates and compares how successful the methods were in predicting personality traits depending on condition "self", "other" and both combined. The p-values show the significance of the difference in correlations across traits.

BIG5	*r*(RS1,RS2)	*r*(W1,W2)	*p*
All	376		
Ext.	0.49	0.55	0.2609
Open.	0.19	0.43	0.0003
Agree.	0.15	0.47	0.0000
Con.	0.22	0.47	0.0001
Neuro.	0.38	0.48	0.0932
Self	215		
Ext.	0.51	0.47	0.5877

Open.	0.12	0.25	0.1651
Agree.	0.20	0.26	0.5141
Con.	0.23	0.25	0.8270
Neuro.	0.31	0.23	0.3740
Others	161		
Ext.	0.43	0.64	0.0080
Open.	0.12	0.50	0.0001
Agree.	0.03	0.60	0.0000
Con.	0.33	0.61	0.0011
Neuro.	0.45	0.65	0.0098

Note: The Big Five traits have been abbreviated as follows; Ext. - *Extraversion*, Open. - *Openness*, Agree. - *Agreeableness*, Con. - *Conscientiousness* and Neuro. - *Neuroticism*.

Word clouds

Word clouds (figures 1-5) were created to visualize the frequency of words used to describe the five different personality traits. The more frequently the words were used by the participants, the larger it appears in the clouds, as well as the centered words in the figures are the most indicative ones.

Figure 1.

Keyword frequency - neuroticism.

Figure 2.

Keyword frequency - conscientiousness

Figure 3.

Keyword frequency - agreeableness.

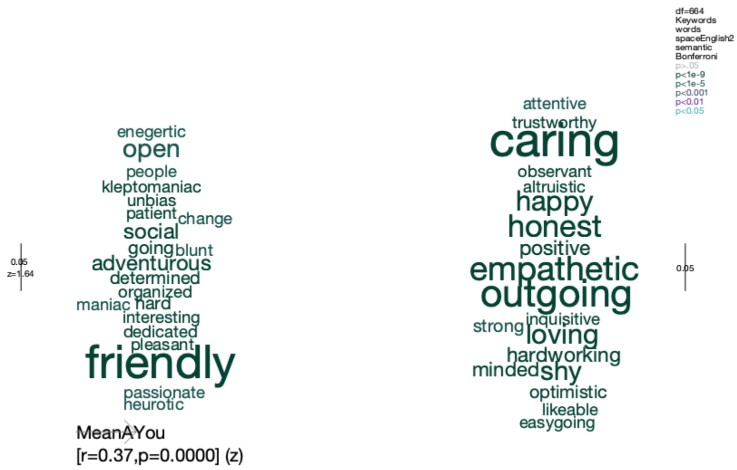


Figure 4.

Keyword frequency - openness.

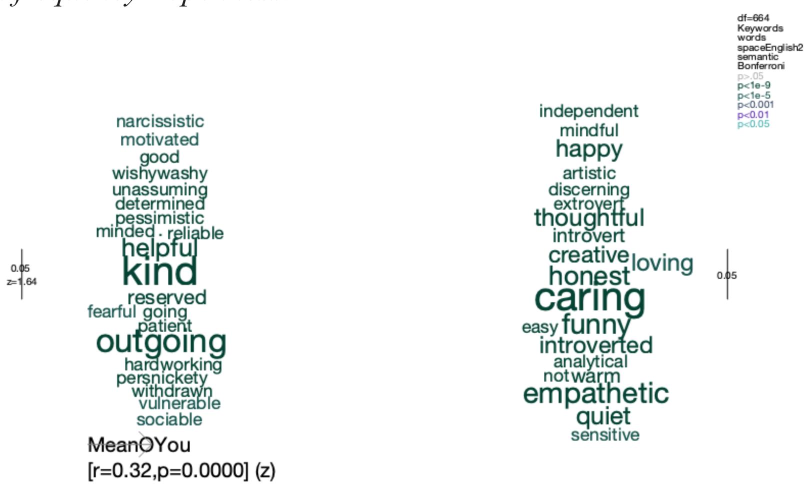
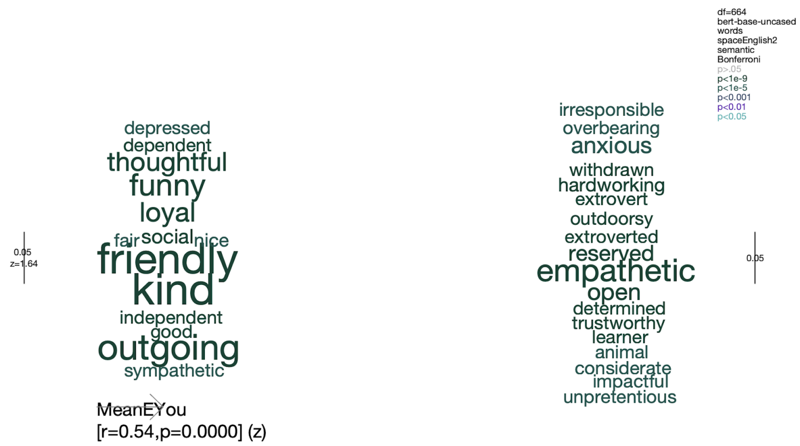


Figure 5.

Keyword frequency - extraversion.



Discussion

This study aimed to investigate the validity of semantic measures compared to rating scales with regards to assessing personality constructs. The results showed that participants' personality narratives were overall more accurately categorized by computational language based responses (semantic responses, text and keywords) compared to rating scale responses.

Support was found for the first hypothesis, namely that semantic measures would prove better at predicting personality constructs compared to rating scales. The results showed that both semantic measures and rating scales combined indeed had a higher validity than either method on its own, however it was not significantly higher than semantic measures on its own, therefore the second hypothesis is rejected.

Validity of Semantic Measures

Rating scales have formerly been believed to have higher validity than language based responses. However the findings in this study, consistent with the findings in other recent studies using computational language assessment, offers evidence that open-ended semantic measures is a viable alternative to investigating personality as it shows higher validity than the widely used rating scale IPIP-NEO. The percentage of correct categorizations of personality traits across both phases combined was 54% when based on semantic responses and 44% when based on the rating scale responses, a difference of 10% between the two methods. The highest score of correct categorizations however, was reached with both methods combined: 54%, slightly better than the semantic method on its own (see table 1).

While the results weakly indicate that semantic measures and rating scales combined are more successful in predicting personality traits overall, it is worth noting that the combination was only slightly better than semantic measures on its own, 1% more correct categorizations and this difference is not statistically significant.

Furthermore, when investigating the correlations for predictions of the personality traits based on semantic responses and rating scales, the accuracy varied greatly depending on both test conditions,“self” or “other”, as well as the targeted personality trait. As shown in Table 2, the semantic measure method was only significantly more accurate in predicting personality traits when analyzing responses from participants in the “other” condition, that is, when analyzing the semantic responses used generated from describing someone else’s personality.

In the “self” condition however, the semantic measure method did not significantly outperform the rating scale in predicting personality traits. This may be the result of a potential condition bias, as the participants in the “other” condition of Phase 1 were primed with characteristics of a personality trait, before being asked to select and describe the whole personality of a person they know. This was not the case for participants in the “self” condition as they were instructed to describe their own personality as a whole, without first being primed with trait characteristics. As such, the priming and overall difference in instruction may have influenced participants' semantic responses despite efforts to keep conditions as similar as possible. This prompts a need for further investigation as the “self” condition is arguably the most interesting to look at. In order for this method to be a viable replacement it has to outperform rating scales in the “self” condition as well. It is not a valid method if it only reliably produces better results when based on narratives that were generated after participants were primed with certain characteristics.

As previously mentioned the correlations vary greatly depending on the targeted trait, for both the semantic measures and the rating scale (see Table 2). Extraversion consistently scores high correlations across conditions suggesting that this trait is easily assessed using both methods. The trait openness consistently scored comparatively low correlations when based on rating scales whereas the correlations are comparatively high when based on the semantic measures. It is important to note that openness has the least amount of data points out of all the traits. Participants that received this trait upon entering the survey dropped out at a higher rate than any of the other traits, as such the other traits have a higher and more evenly distributed number of data points. This is indicative of an attrition bias in the study and has to be taken into consideration when interpreting the results. One potential explanation for the high dropout rate for this trait is that openness may be harder for people to understand, identify and describe. This is in agreement with research stating that the trait openness is less widely understood and more prevalent in some cultures than others. It is also the most recently added trait in the five factor model (Piedmont & Aycock, 2007).

While using semantic measures as a means to investigate personality constructs, it indeed shows promise as an alternative method to use instead of or in conjunction with existing methods. It is important to still consider, highlight and address potential threats to the validity and reliability of the generated results and the method as a whole. Semantic measures could offer more nuance than rating scales, however it is still reliant on written language which may not accurately represent a person's spoken language and may therefore miss aspects of personality which are revealed through speech or body language.

Rating scales, however, offer even less freedom of expression, and consistently perform comparatively poorly when used to predict personality. This is worth highlighting as it is currently the most widely used method.

Furthermore, certain people may have difficulties in expressing themselves through written language, especially in a research context such as this one. Which leads to another factor to consider: language is context sensitive. The ways in which individuals choose to express themselves and respond to questions may vary greatly depending on the context, such as situation or audience. Therefore, the results derived often come from participants' limited written expressions of personality and results should be interpreted with this in mind.

Another point of interest is the possibility of social desirability bias skewing the results. Participants may have tended towards highlighting characteristics that they think are perceived as socially desirable instead of negative ones. As for the “other” condition, participants may not want to be perceived as rude or mean to others by interpreting them and describing them in a manner that could be considered non-flattering. This may be reflected in the word clouds (see Figures 1-5). The word clouds contain mostly positive words, with “caring”, “kind” and “friendly” being the most frequently recurring keywords regardless of targeted trait. Likely this is also a symptom of context sensitivity, as participants could be writing down the first keywords that come to mind when responding to the survey, rather than reflecting deeper upon it. It could also simply be the case that these descriptive words fit into several personality traits respectively, and are not indicative of any of the traits.

Limitations

Limitations and possible sources of errors were detected in the data collection phase. A recurrent issue was that participants did not sufficiently understand the definition of personality. Participants were given a description (see Appendix A) in conjunction with the questions, however misunderstandings still occurred. This could be a result of the provided description not being clear enough. It may also be a result of participants not reading the description attentively or a result of them having a different understanding of what personality entails when answering the questions. It may also be a consequence of the survey being rather cognitively demanding.

A limitation regarding the questionnaires is that the narratives gathered in Phase 1, and subsequently shown to participants in Phase 2, could not be seen as participants completed the survey. Participants had to remember the narratives or go back and forth between question and text. A function for the text to be a continuous part of the whole questionnaire could not be found when constructing the survey. Given this, participants' responses may have been influenced and the rating scale may have been disproportionately affected as it was at the end of the questionnaire and demands more cognitive resources of the participant when answering, compared to the single semantic question, that is writing the keywords.

Implications and Future Research

What then, can be said about the validity of semantic measures contrary to rating scales in assessing personality? First of all, semantic measures have shown an overall more accurate categorization and prediction of personality traits compared to rating scales. However, notably, semantic measures did not significantly outperform the rating scale in predicting personality traits from semantic responses in the “self” condition.

What is clear with the results is that the rating scale as a method on its own, continuously produced lower scores in predicting and categorizing personality. This suggests that semantic measures may be a more updated way of exploring and assessing personality.

Further research should be conducted regarding the applicability of semantic measures in the field of personality psychology, by looking into the possible impact of the condition bias on the results and whether reliable results can be achieved within a research context reminiscent of the “self” condition in this study, in order to strengthen the validity of the method and prove its usefulness in the field.

By refining the procedure used in this study and taking the errors highlighted in this paper into account when conducting future research, new results may very well indicate that semantic measures is a more valid method across both conditions, making it a fully suitable replacement for rating scales.

Under more controlled conditions, such as in a clinical setting, participants can be expected to think more closely about the way they respond to a semantic question as it would be within their own best interest to be understood, in order to receive the most beneficial care. This would likely lead to more accurate predictions and classifications of personality using this method. It can also be argued that the new expressions of personality generated through this method may very well lead to enhanced understanding of both individuals and personality constructs as a whole.

Looking into the possibility of letting people communicate their personalities verbally and then transcribing the semantic responses as a means of using the same semantic measures, while giving participants the opportunity to express themselves in another natural way could be an interesting, although resource demanding way of exploring personality further. Perhaps even comparing the verbal responses to written ones and investigating how language channels affect content and self expression.

Furthermore, looking into the possibility of applying this method to even more fields of psychology could be a valuable way of further investigating the validity of the method while potentially generating new knowledge about the targeted field, as well as possibly updating the way certain constructs are viewed. It would also be of interest to investigate whether other computational methods, such as those involving deep learning neural networks, could be applied to study psychological constructs, using semantic responses in a similar way to SemanticsExcel. Further updating the field of psychology research in correspondence with technological advancement in order to challenge established methods and contribute to new understanding of the human psyche and the ways in which it can be expressed which could benefit people at a large scale.

Lastly, since half of the participants in this study consisted of females, it would be interesting for future research to apply a gender perspective while analyzing the results. Women generally tend to be perceived as more communicative, caring and empathetic in relation to men, a gender perspective on semantic measures versus rating scales might contribute to deeper understanding of possible differences in womens and mens communication styles and perspectives on personality.

Conclusions

In conclusion, the results generated in this study underscore the value of complementary approaches while studying personality, and psychology as a whole. This combined method highlights the promise of valid results without the risk of compromising individual expression and nuance all too much. However, the method needs refining and further investigation as reliable valid results were not reached for the "self" condition. This should be looked into before the method can be reliably applied in the field of personality research. What can be ascertained is the potential semantic measures have within personality psychology research, however rating scales still hold their relevance by producing reliable results and thus should not be denounced.

References

Caprara, G. V., Barbaranelli, C., Borgogni, L., & Perugini, M. (1993). The “big five questionnaire”: A new questionnaire to assess the five factor model. *Personality and Individual Differences*, *15*(3), 281–288. https://doi.org/10.1016/0191-8869(93)90218-r

Costa, P. T., & McCrae, R. R. (1990). Personality Disorders and The Five-Factor Model of Personality. *Journal of Personality Disorders*, *4*(4), 362–371. https://doi.org/10.1521/pedi.1990.4.4.362

Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment*, *4*(1), 5–13. https://doi.org/10.1037/1040-3590.4.1.5

Digman, J. M. (1990). Personality Structure: Emergence of the Five-Factor Model. *Annual Review of Psychology*, *41*(1), 417–440.

Franić, S., Borsboom, D., Dolan, C. V., & Boomsma, D. I. (2013). The Big Five Personality Traits: Psychological Entities or Statistical Constructs? *Behavior Genetics*, *44*(6), 591–604. https://doi.org/10.1007/s10519-013-9625-7

Hinz, A., Michalski, D., Schwarz, R., & Herzberg, P. Y. (2007). The acquiescence effect in responding to a questionnaire. *GMS Psycho-Social Medicine*, *4*, Doc07. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2736523/#:~:text=Acquiescence%20(yes%2Dset)%20describes

Kjell, O. N. E., Kjell, K., Garcia, D., & Sikström, S. (2019). Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods*, *24*(1), 92–115. https://doi.org/10.1037/met0000191

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. https://doi.org/10.1016/j.paid.2016.07.010

McCrae, R. R. (2010). The Place of the FFM in Personality Psychology. *Psychological Inquiry*, *21*(1), 57–64. https://doi.org/10.1080/10478401003648773

McCrae, R. R., & Costa, P. T. (1985). Updating Norman's "adequacy taxonomy": Intelligence and personality dimensions in natural language and in questionnaires.. *Journal of Personality and Social Psychology*, *49*(3), 710–721. https://doi.org/10.1037/0022-3514.49.3.710

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*(1), 81–90. https://doi.org/10.1037/0022-3514.52.1.81

McCrae, R. R., & Costa, P. T. (1997). Personality trait structure as a human universal. *American Psychologist*, *52*(5), 509–516. https://doi.org/10.1037/0003-066x.52.5.509

McCrae, R. R., & Costa, P. T. (2013). Introduction to the empirical and theoretical status of the five-factor model of personality traits. *Personality Disorders and the Five-Factor Model of Personality (3rd Ed.).*, 15–27. https://doi.org/10.1037/13939-002

McDaniel, P., Papernot, N., & Celik, Z. B. (2016). Machine Learning in Adversarial Settings. *IEEE Security & Privacy*, *14*(3), 68–72. https://doi.org/10.1109/msp.2016.51

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, *77*(6), 1296–1312. https://doi.org/10.1037/0022-3514.77.6.1296

Piedmont, R. L., & Aycock, W. (2007). An historical analysis of the lexical emergence of the Big Five personality adjective descriptors. *Personality and Individual Differences*, *42*(6), 1059–1068. https://doi.org/10.1016/j.paid.2006.09.015

Pilehvar, M. T., & Camacho-Collados, J. (2020). Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning. *Synthesis Lectures on Human Language Technologies*, *13*(4), 1–175. https://doi.org/10.2200/s01057ed1v01y202009hlt047

Uher, J. (2018). Quantitative Data From Rating Scales: An Epistemological and Methodological Enquiry. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02599

Yarkoni, T. (2010). Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, *44*(3), 363–373. https://doi.org/10.1016/j.jrp.2010.04.001

Appendix A

Questionnaire Phase 1 - semantic questions

A.1. Text

A.1.1. *Yourself*

Please write a text that describes your personality. Please write a paragraph, approximately five sentences (minimum 300 characters). Write about those aspects that are most important and meaningful to you.

A.1.2. *Agreeableness*

Please think of a person you know well and that matches the personality trait of "agreeableness" (people with this trait are cooperative and empathetic). Take some time to select a specific person that you know (i.e., a friend or acquaintance) and that are high on "agreeableness".

Please write a text that describes **this person's personality as a whole**. Do NOT write a general description of the "agreeableness" trait, instead describe the personality of the person that you selected. Please write a paragraph, approximately five sentences (minimum 300 characters). Write about those aspects that you consider are the most important and meaningful. Note. Please do not use the words "agreeable, cooperative or empathetic" in your text."

A.1.3. *Conscientiousness*

Please think of a person you know well and that matches the personality trait of "conscientiousness" (people with this trait are organized and attentive). Take some time to select a specific person that you know (i.e., a friend or acquaintance) and that are high on "conscientiousness".

Please write a text that describes **this person's personality as a whole**. Do NOT write a general description of the "conscientiousness" trait, instead describe the personality of the person that you selected. Please write a paragraph, approximately five sentences (minimum 300 characters). Write about those aspects that you consider are the most important and meaningful. Note. Please do not use the words "conscientious, organized or attentive" in your text.

A.1.4. *Neuroticism*

Please think of a person you know well and that matches the personality trait of "neuroticism" (people with this trait often experience negative emotions and insecurities). Take some time to select a specific person that you know (i.e., a friend or acquaintance) and that are high on "neuroticism".

Please write a text that describes **this person's personality as a whole**. Do NOT write a general description of the "neuroticism" trait, instead describe the personality of the person that you selected. Please write a paragraph, approximately five sentences (minimum 300 characters). Write about those aspects that you consider are the most important and meaningful. Note. Please do not use the words "neurotic or insecure" in your text.

A.1.5. *Extraversion*

Please think of a person you know well and that matches the personality trait of "extraversion" (people with this trait are sociable and assertive). Take some time to select a specific person that you know (i.e., a friend or acquaintance) and that are high on "extraversion".

Please write a text that describes **this person's personality as a whole**. Do NOT write a general description of the "extraversion" trait, instead describe the personality of the person that you selected. Please write a paragraph, approximately five sentences (minimum 300 characters). Write about those aspects that you consider are the most important and meaningful. Note. Please do not use the words "extrovert, sociable or assertive" in your text.

A.1.6. *Openness*

Please think of a person you know well and that matches the personality trait of "openness" (openness refers to openness to experiences, people with this trait are often curious and imaginative.). Take some time to select a specific person that you know (i.e., a friend or acquaintance) and that are high on "openness".

Please write a text that describes **this person's personality as a whole**. Do NOT write a general description of the "openness" trait, instead describe the personality of the person that you selected. Please write a paragraph, approximately five sentences (minimum 300 characters). Write about those aspects that you consider are the most important and meaningful. Note. Please do not use the words "open, curious or imaginative" in your text.

A.2. Keywords

A.2.1. *Yourself*

Write 5 keywords that best captures your personality

A.2.2. *Agreeableness*

Write 5 keywords that best captures the whole personality of the person you wrote about in the previous question.

Please don't use the word "agreeable"

A.2.3. *Conscientiousness*

Write 5 keywords that best captures the whole personality of the person you wrote about in the previous question.

Please don't use the word "conscientious"

A.2.4. *Neuroticism*

Write 5 keywords that best captures the whole personality of the person you wrote about in the previous question.

Please don't use the word "neurotic"

A.2.5. *Extraversion*

Write 5 keywords that best captures the whole personality of the person you wrote about in the previous question.

Please don't use the word "extroverted"

A.2.6. *Openness*

Write 5 keywords that best captures the whole personality of the person you wrote about in the previous question.

Please don't use the word "open"

Appendix B

Questionnaire Phase 2 - semantic questions

Please read the text below and write 5 keywords that best captures the personality of the author

Free-text answers from participants in Phase 1 were inserted below.