A Comparison of the Assessments from a Person and the Person's AI-Model Using Computational Language Assessments

Eijsbroek, Veerle

A Comparison of the Assessments from a Person and the Person's AI–Model Using Computational Language Assessments

Veerle Eijsbroek
Master thesis
Research Master Clinical and Health Psychology, Leiden University

Daily supervisor: Oscar Kjell, Lund University
First examiner: Eiko Fried, Leiden University

May 26, 2023

Open data and code: https://osf.io/p4xzj/

Words: 7989
(excluding abstracts, tables, figures, acknowledgement, references, and appendices)

Abstract

Across clinical assessment tasks, a statistical model trained on the assessments of one person (a *person's model*) has been shown to be more accurate than the person on which the model is based, the *Model-over-Person* effect. Because the language that people use to express their state of mind is clinically meaningful, the objective of this study was to examine whether the Model-over-Person effect extends to language assessments as well as to identify conditions in which the effect occurs. The accuracy of the assessments of a person versus a person's model was measured as their agreement with a reference standard (the mean assessment of multiple assessors) in two conditions: 1) the assessment of single words and 2) the assessment of texts. Artificial Intelligence based language assessments were employed to create the person's model. No Model-over-Person effect occurred in the assessment of single words or all texts (N = 500 words/texts). A small Model-over-Person effect took place for all three assessors in the assessment of the longer texts ($\geqslant$ 50 words; d_z = .39-.42; n = 23 texts). This effect be explained by the finding that a high amount of input data can make an assessment more prone to human error. Additionally, the relation between the accuracy and different assessment and language characteristics indicated that a person's model could be more accurate in case of a low agreement among assessors and that the accuracy is not related to the confidence of the assessor in the assessment. The results show how computational language assessments can complement a person in accuracy and may support the use of computational language models as decision-support in clinical decision-making.

Layman's abstract

Psychological assessments can be performed by a person or by using a statistical model. A statistical model can be based on the assessments of one person: a *person's model*. The assessments of a person's model can be more accurate than the assessments of the person on which the model is based: the *Model-over-Person* effect.

Because language is the most natural way for someone to express themselves, assessing someone's language is relevant for psychological assessments. The first goal of this study was to examine if the Model-over-Person effect takes place in language assessments. The second goal was to identify situations in which the effect takes place.

The accuracy of a person's model was compared to the accuracy of a person in two situations. The first situation was the assessment of single words. The second situation was the assessment of texts consisting of multiple words. The accuracy of the person and the person's model was measured as their agreement with the average assessment of multiple people. Artificial Intelligence technology was used to create the person's model.

The results showed that the Model-over-Person effect did not take place in the assessment of single words or in the assessment of all texts that varied in length. A small Model-over-Person effect was found in the assessment of the longer texts. An explanation for this finding is that a high amount of information can make it more difficult for a person to assess the information accurately.

Next to these situations, the influence of different assessment and language characteristics was researched. The results showed that a person's model can be more accurate than a person when there is a low agreement among multiple people. Also, the accuracy of the person does not seem to be related to his or her confidence in the assessment.

The findings show how Artificial Intelligence technology can complement a person in assessing language accurately. The findings may support the use of statistical models in psychological assessments.

Introduction

Accurate psychological assessments are crucial for the early detection and prevention of mental disorders as well as for providing personalized treatments (*precision mental health*; DeRubeis, 2019). Two different approaches to decision-making have been identified for psychological assessments: The clinical and the statistical approach (Dawes, Faust, & Meehl, 1989; Meehl, 1954). Clinical assessments are based on human judgment: A person integrates different sources of information based on personal judgment, clinical experience and/or theoretical perspectives to assess the outcome (Dawes et al., 1989; Meehl, 1954). Statistical assessments are based on empirically established relations between the different sources of information and the outcome of interest: A statistical model is used to combine the different sources of information and assess the outcome (Dawes et al., 1989; Meehl, 1954; see Table 1 for key terms and how they are defined in the current study).

Statistical models are commonly trained on a reference assessment, an outcome established as the best-estimate assessment for a certain combination of input information (Dawes et al., 1989; Meehl, 1954). In case a reference assessment is lacking, a statistical model can also be trained on the assessments of one person (Goldberg, 1970), resulting in a *person's model*. The assessments of a person's model have been shown to be more accurate than the assessments of the person on which the model is based (Camerer, 1981; Goldberg, 1970), which will be called the *Model-over-Person* effect.

Because the language that people use to express their state of mind is clinically meaningful (Kjell, Johnsson, & Sikström, 2021; Tausczik & Pennebaker, 2010), the current study will examine whether the Model-over-Person effect extends to language assessments with the aim of identifying conditions in which the effect occurs. To create the person's model, Artificial Intelligence (AI) based language assessments will be employed (Kjell, Giorgi, & Schwartz, 2023).

Table 1

Key Terms and Definitions

Term	Definition
Assessment	The process of judging, weighing, and combining different sources of information (e.g., the outcomes of different measures) and deciding on which outcome is the most accurate for the given combination of input information.
Clinical assessment	A person judges, weighs, and combines different sources of information based on personal judgment, clinical experience and/or theoretical perspectives to assess the clinical outcome of interest.
Statistical assessment	A statistical model, commonly trained on a *reference assessment*, is used for weighing and combining the different sources of information to estimate the clinical outcome of interest.

Reference assessment	An outcome that is established as the best-estimate assessment for a certain combination of input of information, which is used to *train* the statistical model for statistical assessment.
Reference standard	An outcome that is established as the accurate assessment for a certain combination of input information, which is used to *evaluate* the accuracy of the *assessment methods* (e.g., clinical versus statistical assessment).
Person	The human assessor who performs the clinical assessment.
Person's model	A statistical model that is trained on the *assessments of one person* instead of being trained on a reference assessment.
Model-over-Person effect	When the assessments of a *person's model* are more accurate than the assessments of the *person* on which the model is based, evaluated as their agreement with a reference standard.

Note. In the comparisons of clinical versus statistical assessments, the statistical model is commonly trained on a reference assessment (e.g., see the reviews of Ægisdóttir et al. [2006] and Grove et al. [2000]). In the comparison of assessments of a person versus a person's model, the statistical model is trained on the assessments of the person instead of a reference assessment (e.g., see the studies of Camerer [1981] and Goldberg [1970]). The current study evaluates the Model-over-Person effect in the comparison of a person versus a person's model.

Clinical versus Statistical Assessment

Since the 1950s, the accuracy of clinical and statistical assessments has been evaluated as their agreement with a reference standard, an outcome established as the accurate assessment (e.g., a prior diagnosis, an objective or biological marker, or a future behaviour). Meehl (1954) summarized 20 comparative studies and except for one study, statistical assessment resulted in a higher or equal accuracy as clinical assessment. Over the years, comparative studies (e.g., see the review of 100 studies by Dawes et al. [1989]) showed that statistical assessment equalled or exceeded the accuracy of clinical assessment in most cases, which covered different types of assessment tasks in different domains, such as clinical psychology (e.g., predicting mental illness in students; Danet, 1965), psychiatry (e.g., predicting suicide attempts; Gustafson, Greist, Stauss, Erdman, & Laughren, 1977), and medicine (e.g., recommending surgery; Clarke, 1985).

The conclusions favouring statistical assessment were criticized, including that statistical assessment was compared to 'naive clinical decision making' instead of 'sophisticated clinical decision making' (Holt, 1958, 1970; Mann, 1956). The critiques involved issues with 1) the type of assessments (e.g., assessments that assessors were not familiar with), 2) the expertise of the assessors (e.g., assessors with little experience or training), and 3) the data available to the assessors (e.g., too little amount or only quantitative data).

More recently, the accuracy of statistical versus clinical assessment is examined in meta-analyses, which in response to the critiques (Holt, 1958, 1970; Mann, 1956) also investigated the impact of different assessment characteristics (Ægisdóttir et al., 2006; Grove, Zald, Lebow, Snitz, & Nelson, 2000). Of 136 studies in psychology and medicine, 47% favoured statistical assessment, 47% reported an equal accuracy, and 6% favoured clinical assessment. On average, statistical assessments were 10% more accurate (Grove et al., 2000). Of 69 studies on psychological or mental health assessments, 52% favoured statistical assessment, 38% reported an equal accuracy, and 10% favoured clinical assessment. Overall, there was a 13% increase in accuracy for statistical assessment (Ægisdóttir et al., 2006). The differential accuracy was consistent across different types of assessment tasks and domains (Ægisdóttir et al., 2006; Grove et al., 2000). Unexpectedly, the differential accuracy in favour of statistical assessment was larger when the assessors were more familiar with the assessment task or setting (Ægisdóttir et al., 2006). A higher level of training and experience had no impact on the differential accuracy in favour of statistical assessment (Grove et al., 2000), or made it become smaller, resulting in an equal accuracy (Ægisdóttir et al., 2006). When comparing studies with highly reliable reference standards (e.g., an operative verification for brain damage; Heaton, Grant, Anthony, & Lehman, 1981) versus less reliable reference standards (e.g., a supervisor rating for academic success; Kelly & Fiske, 1950), the differential accuracy was consistent for reference standards of varying quality (Ægisdóttir et al., 2006). Finally, when the amount of data available to the assessors increased, the differential accuracy stayed consistent (Grove et al., 2000) or became larger in favour of statistical assessment (Ægisdóttir et al., 2006).

On balance, statistical assessment has been shown to achieve higher accuracy than clinical assessment across different types of assessment tasks in different clinical domains, but findings favouring clinical assessment are present in every domain and certain assessment characteristics have been shown to impact the differential accuracy (Ægisdóttir et al., 2006; Grove et al., 2000). The inconsistent differential accuracy calls for researching the conditions in which each assessment method is the most accurate.

Validity of Statistical versus Clinical Assessments

When a statistical model is trained on a reference assessment using enough training data to learn reliable relations, the statistical properties ensure that the different sources of input information contribute to the assessment based on their established predictive power in their relation to the outcome of interest (Garb & Woord, 2019; Grove & Meehl, 1996). The validity of clinical assessments is dependent on the validity of the assessment strategy (i.e., how the different sources of information are judged, weighted, and combined; Dawes et al., 1989; Grove & Meehl., 1996). Human judgment in clinical decision-making is prone to several biases (Bowes, Ammirati, Costello, Basterfield, & Lilienfeld, 2020; Saposnik, Redelmeier, Ruff, & Tobler, 2016). Examples include the *confirmation bias* (i.e., focussing on information that aligns with your beliefs) and the *anchoring*

heuristic (i.e., being overly influenced by initial information and not sufficiently updating with new information; Bowes et al., 2020). Unconscious systematic biases can result in less valid assessment strategies and less accurate assessments, especially when the amount of information to judge, weigh, and combine is high (Ægisdóttir et al., 2006).

Reliability of Statistical versus Clinical Assessments

The statistical properties of statistical models ensure that the assessment outcome will always be the same for a given combination of input information (Garb & Wood, 2019; Grove & Meehl, 1996). Contrarily, human judgment in clinical decision-making is subject to random error or noise (Dawes & Corrigan, 1971; Grove & Meehl, 1996). Factors such as fatigue, boredom, recent experiences, and distractions can cause random fluctuations that decrease the reliability of the assessments (Dawes & Corrigan, 1971; Grove et al., 2000). Together with a high amount of information to judge and combine, these factors can result in a cognitive overload and a high level of error (Ægisdóttir et al., 2006; Grove & Meehl, 1996). Due to an assessor's unreliability, a valid assessment strategy will not consistently result in an accurate assessment (Dawes & Corrigan, 1971).

Person versus Person's Model

A person's assessment strategy can be separated from the person's random errors by letting their assessments be executed by a statistical model that is trained on the assessments of this person (a *person's model*; Camerer, 1981; Dawes & Corrigan, 1974). Statistical models that are trained on the assessments of one person are known as *paramorphic models* (i.e., representing the assessor statistically) or *bootstrapping models* (i.e., replacing the assessor by their statistical representation) and are useful in cases where a reference assessment for training the model on is lacking (Dawes & Corrigan, 1974). When a statistical model is trained on the assessments of one person, the model has been shown to be more accurate than the assessments of that person (Camerer, 1981; Goldberg, 1970). An explanation for this *Model-over-Person* effect is that professionals can generate accurate assessment strategies, but that statistical models can learn these strategies and then execute them with greater consistency and reliability (Dawes & Corrigan, 1974; Dudycha & Naylor, 1966).

Based on statistical models of 29 clinicians that assessed psychosis versus neurosis diagnoses for 861 patients, the models turned out to be more accurate than (86%) or at least equally accurate as (97%) the clinicians on which they were based (Goldberg, 1970). The Model-over-Person effect occurred for 25 clinicians, including for the most accurate, least accurate, and average assessor. The average difference in correlation of the reference standard (a prior diagnosis) with the clinician's models (r = .31) and with the clinician's assessments (r = .28) was small (.03; Goldberg, 1970). Later, 15 studies were reviewed (Camerer, 1981) that covered different types of assessment tasks in clinical contexts (e.g., predicting intelligence scores of psychiatric patients; Grebstein, 1963), business contexts (e.g., estimating price changes; Wright, 1979) and academic contexts (e.g., predicting

academic success; Wiggins & Kohen, 1971). The person's models were at least as accurate as the person and the Model-over-Person effect occurred in 13 studies (87%; Camerer, 1981). The average difference in correlation of the reference standard with the person's models (r = .39) and with the person's assessments (r = .33) across the studies was again small (.06; Camerer, 1981).

Critiques on the Model-over-Person effects included that the evidence was conflicting and dependent on procedural characteristics (Libby, 1976a, 1976b). When analyzing the same financial dataset to predict business failure/success, some researchers reported the models to be more accurate than the assessors (Goldberg, 1976) and others found the assessors to be more accurate than their models (Libby, 1976a). The conclusion that for different assessors (Goldberg, 1970) and across different assessment tasks in different domains (Camerer, 1981), a person's model is generally more accurate than the person themselves was argued to be overestimated and overgeneralized (Libby, 1976a, 1976b). In summary, to find out when the Model-over-Person effect occurs, the findings should be replicated, and specific conditions should be identified when a person's model surpasses the person in accuracy.

Lack of Impact

Despite the repeated conclusions favouring statistical over clinical assessment, the research had little impact on clinical decision-making procedures for which several reasons have been proposed (Ægisdóttir et al., 2006; Grove & Meehl, 1996). The lack of impact could be due to a lack of familiarity with the scientific evidence, or it could be known but dismissed due to misconceptions, such as assuming biased comparisons. Professionals' education, theoretical orientations and personal values could not align with the evidence and reinforce the resistance. Confirmatory biases and overconfidence of professionals could contribute to the preference for the clinical approach. Finally, psychologists may believe that statistical assessment lowers interpersonal sensitivity or dehumanizes their clients. However, it can allow psychologists to spend more time on gathering information for the assessment as well as tasks for which human skills are uniquely necessary, such as building the client-therapist relation and the treatment itself (Davenport & Kalakota, 2019).

Artificial Intelligence (AI)

Since the introduction of the clinical versus statistical controversy, statistical modelling has undergone big developments from AI techniques (Garb & Wood, 2019; Graham et al., 2020). AI technology has become increasingly prevalent in everyday life as well as in healthcare, especially in medicine (Davenport & Kalakota, 2019). AI technology has been shown to be able to facilitate the early detection of diseases; support the assessment of disease progression; and optimize decisions regarding treatment and medication (Lee et al., 2021). Clinical validation and readiness for implementation should be weighed heavily before employing AI technology since flawed algorithms due to, for example, small or biased training data can have impactful adverse effects (Topol, 2019).

Also, issues such as accountability, transparency, and explainability should receive serious attention (Habli, Lawton, & Porter, 2020; Leslie, 2019).

AI techniques used in health care and mental health care include for example Machine Learning and Natural Language Processing (Graham et al., 2020; Topol, 2019). In Machine Learning, an algorithm is used to combine various sources of data (e.g., the outcomes of different measures) to train a computational model to predict different types of outcomes (e.g., diagnoses or symptoms; Bzdok et al., 2018). In combination with Natural Language Processing, a computational language model can be trained to predict various outcomes (e.g., diagnoses or severity of symptoms) based on quantified text data (e.g., clinical or non-clinical writings; Hirschberg & Manning, 2015). The possibility of quantifying the meaning of text data through Natural Language Processing meets the critique that only quantitative or coded qualitative data can be included in statistical assessments (Holt, 1958, 1970).

Computational Language Models

Since words are the natural medium for people to express their state of mind, people's language contains rich psychological information (Tausczik & Pennebaker, 2010). The language we use has been shown to reflect our emotions (e.g., Sun, Schwartz, Son, Kern, & Vazire, 2020), behaviours (e.g., Kjell, Daukantaitė, & Sikström, 2021), mental health (e.g., Kjell, Johnsson et al., 2021), and personality (e.g., Schwartz, Eichstaedt, Kern et al., 2013). Thanks to Natural Language Processing, people's language can be quantified into psychologically meaningful scores (Kjell, Kjell, & Schwartz, 2023; Schwartz, Eichstaedt, Kern et al., 2013). For example, Natural Language Processing and Machine Learning have been shown to be useful in screening for mental illnesses based on social media language (e.g., Eichstaedt et al., 2018) and Natural Language Processing of question-based language responses have been demonstrated to *measure*, *describe*, and *differentiate* well between psychological constructs (Kjell, Kjell, Garcia, & Sikström, 2019).

Recent advances in Natural Language Processing and Machine Learning, namely *transformers* (Vaswani et al., 2017), have led to increased accuracies across many statistical language processing tasks, such as web search, machine translation, and question answering via chats like ChatGPT (Kjell, Kjell et al., 2023). The increased accuracies can be largely attributed to the capabilities of the models to numerically represent the meaning of words in their context by taking the word context and order into account, which allows for word sense disambiguation (e.g., 'I feel *great*' versus 'I feel like a *great* failure'; Kjell, Giorgi et al., 2023; Vaswani et al., 2017). Transformer-based language models achieved increased accuracies across different human-level Natural Language Processing tasks (i.e., modelling the people behind the language instead of the language itself; Ganesan, Matero, Ravula, Vu, & Schwartz, 2021), such as predicting mental health, personality, and demographics.

Transformer-based language models have been shown to surpass human performances across various human-level and standardized language understanding tasks (Kjell, Kjell et al., 2013). For

example, they surpass the human baseline (i.e., a conservative estimate of human performance from non-expert assessors; Nangia & Bowman, 2019) in different tests from the *General Language Understanding Evaluation* (GLUE; Wang, Singh, Michael, Hill, Levy, & Bowman, 2018). In psychology, transformer-based language models have been demonstrated to perform highly accurate in psychological assessments as validated against rating scales (Kjell, Sikström, Kjell, & Schwartz, 2022) and to perform well in predicting symptoms of depression and anxiety based on question-based language responses (Kjell, Johnsson et al., 2021).

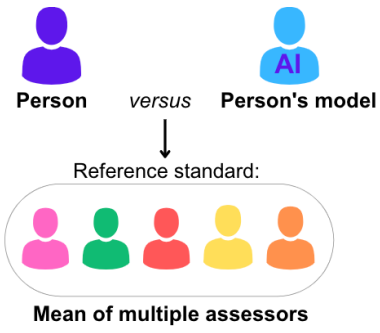
Study

Because the language that people use to express themselves is clinically meaningful and therefore relevant in psychological assessments (Kjell, Johnsson et al., 2021; Tausczik & Pennebaker, 2010), the objective is to examine whether the Model-over-Person effect extends to language assessments. The above conclusions that a person's model generally tends to be more accurate than the person themselves (Camerer, 1981; Goldberg, 1970) call for a specification of the conditions in which a person's model surpasses the person in accuracy. Hence, the aim is to identify conditions in which a person's computational language model is more accurate than the person on which it is based.

In short, the accuracy of the assessments of a person and of a computational language model that is trained on the person's assessments (*the person's model*; Figure 1) will be compared by measuring their agreement with the mean assessment of multiple human assessors (*the reference standard*; Figure 1). To create the person's model, a transformer-based language model will be used (Liu et al., 2019). It is hypothesized that in the assessment of language, the assessments of a person's model are more accurate than the assessments of that person (i.e., the Model-over-Person effect extends to language assessments). This means that the agreement between the person's model and the reference standard is higher than between the person and the reference standard (i.e., the person's model's error is lower than the person's error). To identify conditions in which the Model-over-Person effect occurs, the hypothesis is tested in two conditions. From the first to the second, the assessment task increases in the amount of data that has to be assessed, namely from words to texts.

Figure 1

Overview of the Study Design



Methods

The accuracy of the person and the person's computational language model was investigated over two conditions: the assessment of single words (condition 1) and texts constituting multiple words (condition 2). Additionally, other conditions were explored by researching the error of the person and the person's model in relation to different assessment and language characteristics.

Data

The first condition included the assessment of single words ($N = 2895$) from the *Norms for valence, arousal, and dominance for 13,915 English Lemmas* dataset (Warriner, Kuperman, & Brysbaert, 2013). Randomly, 2895 words were selected from the 13915 words to have the same number of cases as was available in the second condition. The second condition included the assessment of Facebook posts (texts constituting multiple words; $N = 2895$) from the *Modelling Valence and Arousal in Facebook posts* dataset (Preoţiuc-Pietro et al., 2016).

Outcome

The *affective valence* of the text data was assessed. Valence (ranging from *positive affect* to *negative affect*) is one of the two dimensions of the circumplex model of affect, which is often used for describing emotional states (Russell, 1980). Processing and experiencing positive or negative emotions are important characteristics of mental states and disorders as well as crucial components of psychotherapy (American Psychological Association, 2013; Tanana et al., 2021). Valence was in the first condition defined as in Warriner et al. (2013): The pleasantness of the emotions invoked by a word going from *unhappy* to *happy* (Osgood, Suci, & Tannenbaum, 1957). A nine-point scale was used ranging from *completely unhappy* (1) to *completely happy* (9; Warriner et al., 2013). In the second condition, it was defined as in Preoţiuc-Pietro et al. (2016): The polarity of the affective content in a post going from *negative* to *positive* (Russell, 1980). Again a nine-point scale was used ranging from *very negative* (1) to *very positive* (9; Preoţiuc-Pietro et al., 2016).

Assessors

The person that assessed the valence of the words as well as the texts was the author of the current study. She is female, Dutch, 25 years old, and has a bachelor's degree in Psychology. Three other people assessed the valence of 500 texts. These assessors are male, Swedish, 27 or 28 years old, and have a bachelor's or master's degree in Psychology. Regarding the existing assessments, the 1827 assessors of Warriner et al. (2013) are people from the US recruited via the Amazon Mechanical Turk crowdsourcing website and aged between 16 and 87 years. Approximately 60% are female and most have some scientific education. The assessors in Preoţiuc-Pietro et al. (2016) are two people from the US with an education in Psychology.

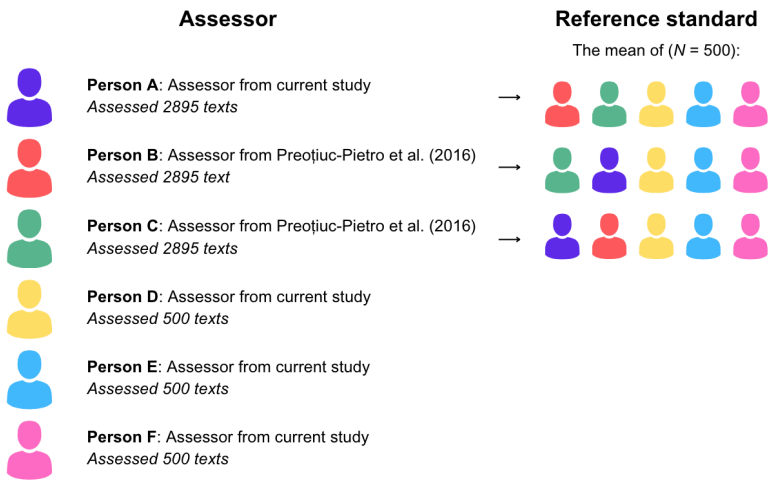
Reference Standard

The mean assessment of multiple assessors was taken as the reference standard because the meaning of language lacks a single objective truth. In the absence of a single objective truth, there is no perfect reference standard that can be used to evaluate the accuracy of the assessment methods. The combination of multiple individual assessments has been proposed as an appropriate solution to attain an acceptable reference standard (Cohen et al., 2016; Bertens et al., 2013). This solution was chosen for the evaluating the language assessments because, since we use language to express ourselves and to communicate with each other, its emotional meaning is based on a shared human understanding (Pennebaker, Mehl, & Niederhoffer, 2003; Tausczik & Pennebaker, 2010). The mean assessment as the reference standard is supported by the finding that the most accurate assessment (when a perfect reference standard is lacking) is the composite judgment of the total group (i.e., the mean assessment of all assessors; Goldberg, 1970).

The person's assessments were not part of the mean assessment that was used as the reference standard to evaluate their and their model's accuracy (Bertens et al., 2013). The reference standard in the first condition was the mean assessment from the assessors of Warriner et al. (2013) based on approximately 20 individual assessments per word. The reference standard in the second condition was the mean assessment of the assessors excluding the person whose assessments were compared to the reference standard (Figure 2). Because the study of Preoţiuc-Pietro et al. (2016) included only two assessors, three more people assessed 500 texts in order to achieve a more reliable mean.

Figure 2

The Assessors and Reference Standards in the Second Condition



Procedure

The assessors provided informed consent before completing their assessments. The words and texts were assessed in a randomized order using an online rating sheet. The definition and operationalisation of valence and the instructions for the assessment task (Appendix A) were similar to the corresponding study from which the assessments were used in the reference standard. The

assessors were blind to the assessments of the other assessors as well as to the assessments from Warriner et al. (2013) and Preoţiuc-Pietro et al. (2016). In the first condition, the person also rated their confidence in each assessment with a percentage between 50 and 100.

Ethics

The current study was conducted in Sweden (Lund University) and was deemed exempt from requiring ethical approval. Swedish law (2003:460) and the Swedish Ethical Review Authority state that only research that 1) includes collecting personal information; and/or 2) involves risk for physical or psychological harm; or 3) involves manipulating or deceiving individuals, should undergo an external ethical review.

Regarding the existing assessments, Warriner et al. (2013) and Preoţiuc-Pietro et al. (2016) provided their text data and valence assessments as open data. The authors of the Facebook posts explicitly gave permission to include their data in a corpus for research purposes after being anonymized by the authors (Preoţiuc-Pietro et al., 2016).

Statistical Analyses

The statistical analyses were performed in R (R Core Team, 2022) using the *text*-package (version 0.9.99.7, https://www.r-text.org/; Kjell, Giorgi et al., 2023), which provides the transformer-based language analysis techniques tailored for social and behavioural scientists. The analyses consisted of 1) applying numeric representations (*word embeddings*) from a pre-trained language model to quantify the words and texts, 2) training these word embeddings to predict the person's assessments of the words and texts, and 3) measuring the accuracy of the person and the person's model as their agreement with the reference standard. The alpha level was .05 in all analyses (two-tailed $p < .05$). The inter-rater reliability in the assessments of the texts was calculated using the Intraclass Correlation (ICC; Shrout & Fleiss, 1979). ICC values less than .50 were interpreted as poor reliability, .50-.75 as moderate, .75-.90 as good, and greater than .90 as excellent (Koo & Li, 2016).

Pre-trained Word Embeddings

The text data were transformed (i.e., quantified) into *word embeddings* (numerical presentations of words). The pre-trained language model that was used is the RoBERTa Large model (Liu et al., 2019). This model is an extended and improved version of the most widely used Bidirectional Encoder Representations from Transformers (BERT; Devlin, Chang, Lee, & Toutanova, 2019) and is trained on over 160 GB of English text including unpublished books, Wikipedia pages, news articles, blogs and stories with use of the Masked Language Model and Next Sentence Prediction objectives (Liu et al. 2019).

An advantage of transformer-based language models in comparison to other language analysis techniques is that they are able to numerically represent words differently according to the context they are in (Kjell, Giorgi et al., 2023; Vaswani et al., 2017). Using the RoBERTa Large model, the

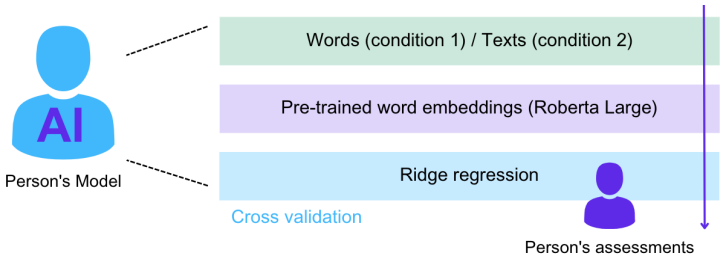
text data were transformed into *bidirectional contextual word embeddings*, which are influenced by the previous and following words in the text. Contextual word embeddings are lists of values (ordered vectors) that numerically represent the meaning of words and capture the relationships between words. The numerical values of vector embeddings can be seen as coordinates in a geometric space comprising hundreds of dimensions. When words are close to each other in this space (i.e., they have similar vector embeddings), they are typically similar in meaning (Kjell, Giorgi et al., 2023). The RoBERTa Large model represents word tokens using 24 layers that comprise 1024 dimensions each (Liu et al., 2019).

Training the Word Embeddings to the Person's Assessments

The data were randomly split into a training set (N = 2395 words/texts; 83%) and a test set (N = 500 words/texts; 17%). To examine the relationship between the words/texts and the valence assessments in the training set (i.e., to train the person's computational language model; Figure 3), the word embedding dimensions of the data were used as predictors in ridge regression to predict the person's assessments. The training was conducted using tenfold cross-validation in which the training set was for the 10 folds repeatedly split into an analysis set (i.e., creating models with different penalties), assessment set (i.e., evaluating the different models), and test set (i.e., applying the best-evaluated model). The search grid for the penalty in ridge regression ranged from 10^{-16}-10^{16}, and the prediction accuracies of the different models were evaluated with Pearson correlations (r) between the observed and the predicted scores.

Figure 3

The Analysis Pipeline to Create the Person's Computational Language Model



Measuring the Accuracy

The person's model was used to predict the valence of the word embedding dimensions in the test set (N = 500). To measure the accuracy of the person versus the person's model, their assessments were correlated to the mean assessment of multiple assessors (the reference standard). Correlations of .20-.39 were interpreted as weak, .40-.59 as strong, .60-.79 as strong and .80-1.00 as very strong (Evans, 1996). Spearman rank correlations (r_s) were used since the assumptions for the parametric Pearson correlations (r) were not met. The error of the person and the person's model was computed by taking the absolute difference between the reference standard and their assessments. Paired sample

t-tests were performed to test whether the absolute errors significantly differed. The valence scores of the person, the person's model, and the reference standard were *z*-transformed before conducting the paired-sample *t*-tests because of differences in distribution and range. Cohen's *d* larger than .20 was interpreted as a small effect, larger than .50 as medium, and larger than .80 as large (Cohen, 1988).

Exploratory Analyses

The relation between the error of the person and the person's model and six assessment and language characteristics was explored. The errors were correlated to the characteristics using Spearman rank correlations (r_z) since the assumptions for the parametric Pearson correlations (r) were not met. To test the differences in the correlation with the person's error and the person's model's error, *z*-difference statistics (z_{diff}) were calculated (Lee & Preacher, 2013).

The investigated characteristics included 1) the valence (*negative* to *positive*) of the mean valence assessments (reference standard scores) and 2) the valence strength (*weak* to *strong*) of the mean valence assessments. To indicate the valence strength, the words and texts were coded on a five-point scale from one (weakly valenced) to five (strongly valenced). Words/texts with a mean assessment of five (*neutral*) were coded as one and words/texts with a mean assessment of one (*very negative*) or nine (*very positive*) were coded as five.

Another characteristic was 3) the variability among the assessors (standard deviations) in the mean valence assessments. The person whose error or model's error was subject to the correlation was not included in the variability estimate (*SD*). The other characteristics included 4) the confidence of the person in the assessments of the words; 5) the number of meanings per word; and 6) the frequency of use per word. To indicate the number of meanings of the words, the lexical database WordNet (Fellbaum, 1998) was used. To determine the frequency of use of the words, the frequency norms from the SUBTLEXUS corpus (Brysbaert & New, 2009) were used.

Results

Descriptive statistics

Person A assessed both the words and texts with a median valence score of five (see Table 2 for the descriptive statistics of the test set; $N = 500$). For example, the word *clock* was assessed with five, *misfortune* with one, and *phenomenal* with nine. The text '*Making toast at 9:17pm'* was, for example, assessed with five, '*My father once told me we are always dying'* with one; and '*Happy Happy Happy !!!!'* with nine. The inter-rater reliabilities among the three assessors who assessed all texts (*ICC* = .74, *95% CI* = .73-.75, $N = 2895$) and among the six assessors who assessed the test set (*ICC* = .75, *95% CI* = .72-.78, $N = 500$) were moderate to good. Warriner et al (2013) reported the split-half reliability ($r = .91$, $N = 13915$) for the assessment of the words since each word was assessed by a different combination of assessors.

Table 2

Descriptive Statistics for the Assessments of the Words and Texts

Assessment and language characteristics	Words (condition 1)		Texts (condition 2)	
	Mean (SD)	Median (Min - Max)	Mean (SD)	Median (Min - Max)
Valence person	5.15 (1.53)	5 (1 - 9)	5.33 (1.81)	5 (1 - 9)
Valence person's model	5.23 (0.86)	5.26 (1.91 - 7.54)	5.30 (1.43)	5.16 (0.97 - 9.47)
Valence reference standard (*M*) [a]	4.98 (1.31)	5.19 (1.62 - 7.94)	5.32 (1.41)	5.40 (1.60 - 8.60)
Number of assessors [b]	23.16 (38.60)	20 (17 - 872)	5	5 (5 - 5)
Variability among assessors (*SD*) [b]	1.66 (0.33)	1.66 (0.68 - 2.60)	0.72 (0.40)	0.71 (0 - 2.12)
Valence strength	1.97 (0.75)	2 (1 - 4)	2.16 (0.88)	2 (1 - 5)
Confidence person (%)	75.60 (16.28)	80 (50 - 100)	N.A.	N.A.
Number of meanings	3.67 (4.91)	2 (1 - 45)	N.A.	N.A.
Frequency of use	1710 (14800)	74 (20 - 314232)	N.A.	N.A.

Note. $N = 500$. The statistics are reported for the person who assessed the words as well as the texts (Person A). The descriptive statistics for the assessments of the texts including Person B to F can be found in Appendix B.

[a] In condition 1, the mean assessment (*M*) of the assessors in Warriner et al. (2013); in condition 2, the mean assessment (*M*) of the assessors excluding the person who's assessments are compared to the reference standard.

[b] The number of assessors on which the reference standard is based and the variability among their assessments (*SD*).

Accuracy of the Person versus the Person's Model

The accuracy of the person's and the person's model was measured as their agreement with the mean assessment of multiple assessors (the reference standard). The correlations with the reference standard of the person and the person's model (Table 3) were strong for the assessment of the words (condition 1), $r_z(498) = .64\text{-}.69$, $p < .001$, and very strong for the assessment of the texts (condition 2), $r_z(498) = .83\text{-}.89$, $p < .001$. Paired-sample *t*-tests were performed to test whether the absolute errors of the person and the person's model (Table 3) significantly differed. No Model-over-Person effect occurred in the assessment of the words or all texts ($N = 500$; Figure 4). However, a small Model-over-Person effect took place for all assessors when examining the longest texts ($\geqslant 50$ words; $d_z = .39\text{-}.42$; $n = 23$; Figure 5).

Words

The person's assessments of the words correlated stronger with the reference standard than their model's assessments in the test set (Figure 4). The paired sample *t*-test showed that the person's error

was significantly lower than the person's model's error, $t(499) = -2.39$, $p = .017$, $d_z = .14$ (Table 3). Therefore, the hypothesis that the person's model is more accurate than the person is rejected for the assessment of single words (i.e., the Model-over-Person effect does not occur).

Table 3

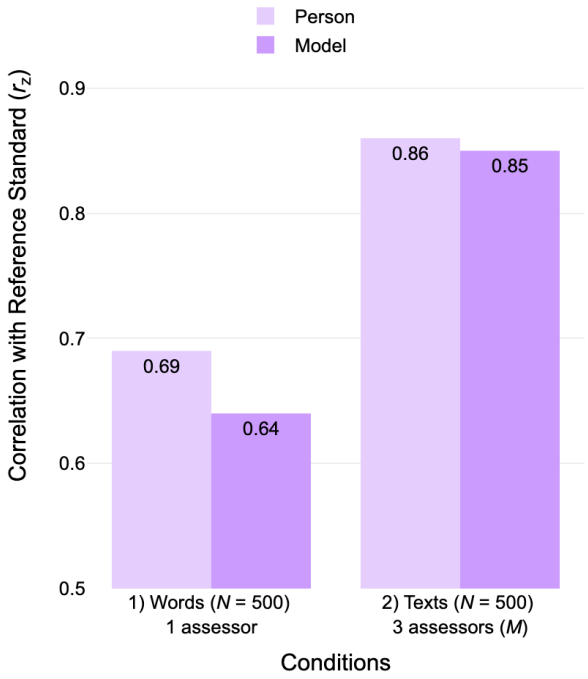
The Correlations with the Reference Standard and the Absolute Errors

Assessor (condition)	Correlation with reference standard r_z		Absolute error Mean$_z$ (SD$_z$)	
	Person	Person's Model	Person	Person's Model
Person A (words)	.69***	.64***	.58 (.48)	.65 (.55)
Person A (texts)	.89***	.84***	.36 (.34)	.48 (.38)
Person B (texts)	.86***	.84***	.43 (.35)	.47 (.38)
Person C (texts)	.83***	.86***	.44 (.39)	.44 (.36)

Note. $N = 500$, *** $p < .001$ (two-tailed).

Figure 4

Correlations with the Reference Standard in the Assessment of the Words and Texts



Note. The training of the person's model in the assessment of the words resulted in a strong correlation between the predicted and observed scores in the training set, $r(2393) = .63$, $p < .001$. For the assessment of the texts, the mean (M) correlation with the reference standard of the three assessors (Person A, B and C) and their models is visualized (see Figure 5 for the correlations with the reference standard per person).

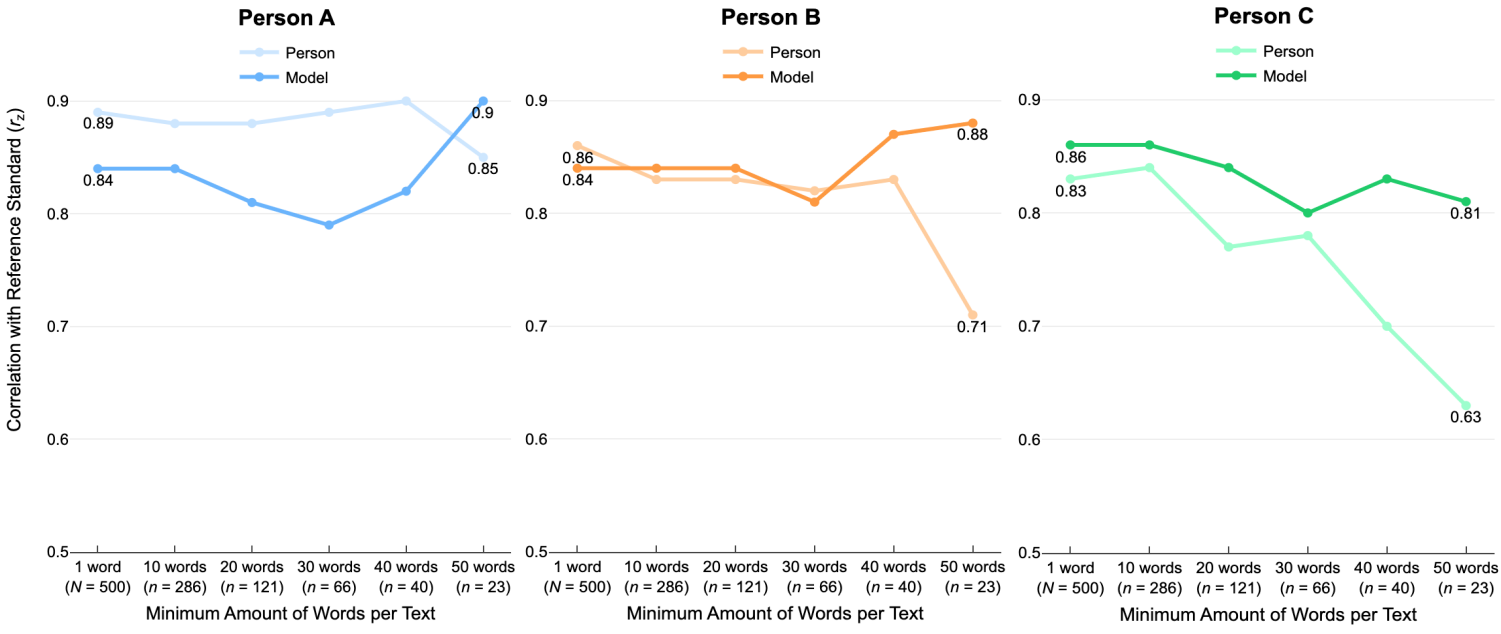
Texts

As for the words, the agreement with the reference standard was on average higher for the person than for the person's model, but with a notably smaller difference (Figure 4). Three persons' models were created because the 2895 texts were assessed by three assessors. For person A, the person's assessments correlated stronger with the reference standard than their model's assessments. Also, the person's error was significantly lower than the person's model's error, $t(499) = -6.57$, $p < .001$, $d_z = .34$. For person B, the person's assessments correlated again stronger with the reference standard than their model's assessments, but the difference in error was not significant, $t(499) = -1.87$, $p = .062$, $d_z = .11$. Contrarily, for person C, the person's model's assessments correlated stronger with the reference standard than the person's assessments, but there was no difference in error, $t(499) = -0.03$, $p = .977$, $d_z = .00$.

From assessing single words to texts, the person's models approached the accuracy of the persons (Figure 4). Therefore, it was decided to look into the assessment accuracy for different minimum amounts of words per text. Figure 5 visualizes the correlations with the reference standard ranging from texts with a minimum of one word per text (all texts; $N = 500$) to texts with a minimum of 50 words per text ($n = 23$).

Figure 5

Correlations with the Reference Standard per Minimum Amount of Words per Text



Note. The training of the person's models in the assessment of the texts resulted in strong correlations between the predicted and observed scores in the training set: $r(2393) = .79$, $p < .001$ for person A; $r(2393) = .71$, $p < .001$ for person B; and $r(2393) = .70$, $p < .001$ for person C.

Only for the longest texts, the hypothesis that the person's model is more accurate than the person holds for all three assessors (i.e., the Model-over-Person effect occurs). For texts with 50 words or more (n = 23, M = 72.65 words, *Median* = 64 words, *Max* = 147 words), the person's model's assessments correlated stronger with the reference standard than the person's assessments. For all assessors, the person's model's error was lower (M_z = .33, M_z = .40 and M_z = .42 respectively) than the person's error (M_z = .47, M_z = .56 and M_z = .64 respectively), but not significantly, $t(22)$ = 1.93, p = .066, d_z = .39; $t(22)$ = 1.62, p = .120, d_z = .42; and M_{diff} = .22; $t(22)$ = 1.23, p = .232; d_z = .39 respectively. The inter-rater reliability of the six assessors in assessing texts with 50 words or more was notably lower (*ICC* = .53, *95% CI* = .36-.72, n = 23) than for all texts (*ICC* = .75, *95% CI* = .72-.78, N = 500).

Assessment and Language Characteristics

The errors of the person and their model were negatively related to the valence (*positive* to *negative*) of the words; positively related to the valence strength (*weak* to *strong*) of the words and texts; and negatively related to the variability among assessors (*SD*) in the assessment of texts (N = 500; Table 4).

Table 4

Correlations between the Assessment and Language Characteristics and the Absolute Errors

Assessment and Language Characteristics	Words (condition 1)		Texts (condition 2)		Texts (condition 2)		Texts (condition 2)	
	Person A		Person A		Person B		Person C	
	Error person	Error model	Error person	Error model	Error person	Error model	Error person	Error model
Valence (*negative - positive*)	-21**	-.32**	.05	.06	.07	-.01	.10*	.03
Valence strength (*weak - strong*)	.24**	.35**	.16**	.13**	.44**	.61**	.32**	.36**
Variability among assessors (*SD*)	.04	-.04	.28**	.15**	.24**	.17**	.28**	.19**
Confidence person (%)	-.04	-.03	-	-	-	-	-	-
Number of meanings	-.11*	-.01	-	-	-	-	-	-
Frequency of use	-.04	.06	-	-	-	-	-	-

Note. N = 500, * p < .05, ** p < .01 (two-tailed). The confidence scores of the person, the number of meanings, and the frequency of use characteristics were not collected or computed for the texts (condition 2).

Valence was negatively related to the person's as well as to the person's model's error in the assessment of words. So, a higher valence (more *positive* words) was associated with less error, and this relation was stronger for the person's model's performance, z_{diff} = 2.13, p = .034. The valence

strength was positively related to the person's as well as to the person's model's error in both conditions. So, a stronger valence (very *negative* or very *positive* words/texts) was associated with more error in the assessment of words and texts. For persons B and C, this relation was stronger for the person's model's performance, z_{diff} = -2.14, p = .033 and z_{diff} = -4.85, p < .001.

The variability among assessors was positively related to the person's and person's model's error in the assessment of texts. So, a higher variability among assessors (i.e., a higher standard deviation in the mean assessment) was associated with more error in the assessment of texts. This relation was stronger for the person's performance, and for person A this difference in correlation was significant, z_{diff} = 2.53, p = .011.

Finally, the number of meanings per word was negatively related to the person's error, and the frequency of use per word was not significantly related to the person's or their model's error.

Discussion

The objective was to examine whether the Model-over-Person effect extends to language assessments and to identify conditions in which a person's computational language model is more accurate than the person. The accuracy of the person versus the person's model was measured for 1) the assessment of words and 2) the assessment of texts. The Model-over-Person effect took place for all assessors when examining the longer texts. The effects were small and non-significant, and the Model-over-Person effect did not occur in the assessment of the words or all texts.

High Amount of Data: Words versus Texts

The Model-over-Person effect occurred for assessing the longer texts and not for assessing single words, which corresponds to previous findings that a higher amount of data available to the assessors increases the relative higher accuracy of statistical assessments compared to clinical assessments (Ægisdóttir et al., 2006). The human brain has limited resources to take into account large amounts of data for inferential purposes (Bowes et al., 2020; Grove & Meehl, 1996): "The human brain is a relatively inefficient device for noticing, selecting, categorizing, recording, retaining, retrieving, and manipulating information for inferential purposes." (Grove & Meehl, 1996, *p.* 23). This is the case in clinical decision-making as well as in everyday life: "Surely we all know that the human brain is poor at weighting and computing. When you check out at a supermarket, you don't eyeball the heap of purchases and say to the clerk, 'Well it looks to me as if it's about $17.00 worth; what do you think?' The clerk adds it up." (Meehl, 1986, *p.* 372). A high amount of data to judge, weigh, and combine can result in a cognitive overload and can make assessments more prone to human errors, which can adversely impact the reliability and accuracy of a person's assessment (Ægisdóttir et al., 2006).

Statistical assessment is not prone to these adverse effects because of two reasons. Firstly, the statistical properties in statistical models ensure that the same combination of input information will

always lead to the same assessment outcome (Garb & Wood, 2019; Grove & Meehl, 1996). Secondly, when the model is trained on enough training data to learn reliable connections and ignore unreliable errors, the statistical properties ensure that the different sources of information will contribute to the assessment based on their reliably established predictive power in relation to the outcome of interest (Grove & Meehl, 1996; Grove et al., 2000). These characteristics of statistical models together with the limited cognitive resources of the human brain can explain why the Model-over-Person effect occurred in the assessment of the longer texts.

Small Model-over-Person Effect

The found Model-over-Person effects were small and non-significant. The small differences in accuracy between the person and the person's model correspond to the previous studies: An average difference in accuracy in favour of the person's models of .03 across the 29 assessors in Goldberg (1970) and an average difference in accuracy of .06 in favour of the person's models across the 15 studies in Camerer (1981). Also, meta-analyses that compared clinical to statistical assessments found an equal accuracy in approximately half of the studies (Ægisdóttir et al., 2006; Grove et al., 2000).

Is a small Model-over-Person effect or an equal accuracy enough to argue that persons could be substituted by their models in clinical assessment tasks? An advantage of the use of statistical assessment (e.g., person's models) in clinical decision-making is that it is faster and can therefore save clinicians' time as well as expenses (Dawes et al., 1989; Grove et al., 2000). Saved time and expenses could be used for other important tasks, such as gathering data for the assessment, building a client-therapist relationship, and the treatment itself. In the case of large assessment tasks, it might be difficult to find a person to do the assessments, whereas it would be possible with computational resources such as a person's model. Also, in some situations, efficiency or consistency might be of greater importance than in other situations. So, the identified differential accuracy can guide the choice for assessments by a person or a person's model, but different situations have different demands, which makes the choice for clinical versus statistical assessment also context-dependent.

Assessment and Language Characteristics

The accuracy of the person's versus their model's assessments was related to the valence (*positive* to *negative*) in the assessment of words; to the valence strength (*weak* to *strong*) in the assessment of words and texts; and to the variability among assessors in the assessment of texts. The errors were negatively related to the valence, so both the person and the person's model performed more accurately in assessing positively than negatively valenced words. The errors were positively related to the valence strength, so both the person and the person's model performed more accurately in assessing weakly than strongly valenced words and texts.

The errors were positively related to the variability among assessors. So, both the person and the person's model performed more accurately in assessing texts with a low than a high variability

among assessors. This relation was stronger for the person's performance, which could indicate that in case of a relatively low agreement among assessors, a person's model could be more accurate than the person. This relation corresponds to the Model-over-Person effect taking place in the assessments of the longer texts in which the inter-reliability of the assessors was lower than in the assessment of all texts. This is supported by the finding that a low inter-rater agreement is an indicator of a relatively low assessment accuracy (Alavi, Biros, Clearly, 2022; McHugh, 2012). So, a low agreement among assessors could indicate that it is harder for a person to assess a case accurately and that it could be better to use the person's model instead of the person's assessment.

The errors were not related to the confidence of the person, which corresponds to the association between confidence and assessment accuracy being mostly small (e.g., $r = .15$ in a meta-analysis including 36 studies and 1485 clinicians; Miller, Spengler, & Spengler, 2015). Confidence has been shown to be a poor indicator of assessment accuracy, and overconfidence has been demonstrated to contribute to errors in different areas of clinical practice including diagnosis and assessment (Miller et al., 2015; Saposnik et al., 2016). So, the choice between the use of statistical (e.g., a person's model) or clinical assessment can better be based on the conditions in which each has been identified as the most accurate than on the assessor's confidence in their assessment.

The errors were not related to the frequency of use per word, and the person's error was negatively related to the number of meanings per word.

Individual Differences

The relative accuracy of the person and the person's model showed individual differences across the three assessors. For one assessor, the Model-over-Person effect only took place for texts with 50 words or more, while for the other assessors the effect already seemed to appear for a lower minimum amount of words per text. Also, the relation of the person's and model's error to certain assessment and language characteristics varied across assessors.

Next to generalisable conditions in which the Model-over-Person effect occurs, there could be conditions that are specific to a certain individual. On an individual level, a person's model could potentially be used to get insight into and educate assessors about their individual strengths and weaknesses: The conditions in which your model surpasses you in accuracy could indicate the conditions in which you are less reliable or accurate and should pay extra attention or decide to use your statistical model. Insight into accuracy and error has been shown to be essential for increasing assessment accuracy in clinical settings but is often lacking (Omron, Kotwal, Garibaldi, & Newman-Toker, 2018; Schiff, 2008). Persons' models could potentially contribute to filling this gap, and future research could examine how these individual conditions relate to characteristics of the assessors, such as their mental states and personality traits, to study the impact of those characteristics in clinical decision-making.

Clinical Relevance

Since the language that people use to express themselves is clinically meaningful (Tausczik & Pennebaker, 2010), assessing language is highly relevant for psychological assessments (Kjell, Johnsson et al., 2021). However, despite the developments and promising results of statistical modelling and language analyses, computational language models are not, to a large extent, employed yet in psychological assessments (Kjell, Kjell et al., 2023). This can be due to factors such as confirmatory biases and overconfidence of professionals that can contribute to the resistance towards statistical assessment methods (Ægisdóttir et al., 2006; Grove & Meehl, 1996). Showing a clinician how computational language assessments can complement them in accuracy by showing them when their own model surpasses them in accuracy (e.g., in the case of long texts) and when not (e.g., in the case of single words) could potentially lower this resistance. Therefore, knowing the conditions in which the Model-over-Person effect occurs could have implications for the adoption of computational language models as a decision-support in psychological assessments.

Limitations and Future Research

Limitations of the current study include the low amount of long texts. In comparison to all texts, the difference in accuracy in favour of the person's model was notably larger for the 23 texts with 50 words or more. More comparisons of persons and their models in assessing long texts are needed to investigate the replicability of the found Model-over-Person effects and to see if the effect becomes larger for longer texts (e.g., more than 100 words).

Another limitation is the low number of assessors. The first condition included one and the second condition three assessors, while previous studies researched the effect across more assessors (e.g., 29 assessors; Goldberg, 1970). Analyses including more assessors are needed to investigate if the effects are consistent across more comparisons between persons and their models and to enable studying individual differences more reliably. Furthermore, the number of assessors included in the mean assessment (reference standard) for the second condition (five) was relatively low as compared to the first condition (approximately 20). The inter-reliabilities in the reference standards were moderate to good when taking all texts into account, but notably poorer for the longer texts. Future research could include more assessors to obtain a more reliable reference standard (DeBruine & Jones, 2018), however, the differential accuracy of statistical versus clinical assessments has been shown to be consistent for reference standards of varying quality (Ægisdóttir et al., 2006).

Regarding the generalisability to clinical settings, future research could investigate if the Model-over-Person effect extends to assessments using clinical text data and outcomes. Assessing valence and assessing (social media) language are both of clinical relevance (e.g., Eichstaedt et al., 2018; Tanana et al., 2021), but future research on person's models could for example use question-based mental health descriptions and focus on assessing diagnoses or symptoms.

Finally, like previous studies on clinical versus statistical assessment, this study has focused on *accuracy*, which can guide which assessment method can be used in which condition. Not choosing the assessment method that has been shown to be the most accurate (clinical or statistical assessment) has even been argued to be unethical, especially in clinical domains in which inaccurate assessments or misclassifications can have impactful adverse effects on people's lives (Dawes et al., 1989; Grove et al., 2000). However, other criteria than accuracy can play a role in which method is the most ethical or appropriate to use, especially in the case of AI techniques (Davenport & Kalakota, 2019; Habli et al., 2020). Criteria such as transparency, explainability, and accountability (Leslie, 2019) are not taken into account and are recommended to receive more attention in future research on statistical versus clinical assessment.

Conclusion

This study complements the past research (Camerer, 1981; Goldberg, 1970) by extending the Model-over-Person effect to a new domain (i.e., language assessments) and identifying conditions in which the effect occurs. The results show that a person's model surpasses the person in accuracy when the amount of text data to judge, weigh, and combine is high. Besides this, the results indicate that a person's model is more accurate than the person when there is a low agreement among assessors and that the decision for a person's model or a person's assessment can better be based on the identified conditions in which each is the most accurate than on the confidence of the assessor.

The findings show how computational language assessments can complement a person in accuracy and may support the use of computational language models as decision-support in clinical decision-making. A person's model could, for example, be used to save a clinician's time or to give clinicians insight into their personal strengths and weaknesses. Therefore, as proposed in the potential future of *precision mental health* (DeRubeis, 2019), statistical assessment methods such as computational language models and person's models might be able to improve accuracy and efficiency in mental health care.

Acknowledgement

I want to thank my supervisor Eiko for his attentiveness, useful feedback, and inspiring collaborations. I want to thank my teacher Marc for his continuous support during my bachelor's and master's programmes. I want to thank my friends Jill, Andrea, Laura, and Annick for making studying in the library fun. I want to thank my mom, dad, and sister for their support and interesting research and healthcare discussions at home. I want to thank the Ablemind team and 'harmony research lab' in Malmö (August and Oskar in special) for providing the most motivating and fun work environment there can be. Finally, I especially want to thank my supervisor Oscar for his optimal guidance, inspiring ideas, and great life and work opportunities.

The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Chalmers University of Technology partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., & Rush, J. D. (2006). The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. *The Counseling Psychologist, 34*(3), 341–382.

Alavi, M., Biros, E., & Cleary, M. (2022). A primer of inter-rater reliability in clinical measurement studies: Pros and pitfalls. *Journal of clinical nursing*, *31*(23-24), e39–e42.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (Vol. 5). American psychiatric association, Washington, DC.

Bertens, L. C., Broekhuizen, B. D., Naaktgeboren, C. A., Rutten, F. H., Hoes, A. W., van Mourik, Y., ... & Reitsma, J. B. (2013). Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. *PLoS medicine, 10*(10), e1001531.

Bowes, S. M., Ammirati, R. J., Costello, T. H., Basterfield, C., & Lilienfeld, S. O. (2020). Cognitive biases, heuristics, and logical fallacies in clinical practice: A brief field guide for practicing clinicians and supervisors. *Professional Psychology: Research and Practice, 51*(5), 435.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods, 41*(4), 977-990.

Bzdok, D., Krzywinski, M., & Altman, N. (2018). Machine learning: supervised methods. *Nature methods, 15*(1), 5.

Camerer, C. (1981). General conditions for the success of bootstrapping models. *Organizational Behavior and Human Performance*, *27*(3), 411-422.

Clarke, J. R. (1985). A comparison of decision analysis and second opinions for surgical decisions. *Archives of Surgery, 120*, 844-847.

Cohen, J. F., Korevaar, D. A., Altman, D. G., Bruns, D. E., Gatsonis, C. A., Hooft, L., ... & Bossuyt, P. M. (2016). STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ open, 6*(11), e012799.

Danet, B. N. (1965). Prediction of mental illness in college students on the basis of "non psychiatric" MMPI profiles. *Journal of Counseling Psychology, 29*, 577-580.

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future healthcare journal*, *6*(2), 94–98.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95-106.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical Versus Actuarial Judgment. *Science*, *243*(4899), 1668–1674.

DeBruine, L. & Jones, B. C. (2018). Determining the number of raters or reliable mean ratings. *Open Science Framework*. Retrieved from: https://debruine.github.io/post/how-many-raters/

DeRubeis, R. J. (2019). The history, current status, and possible future of precision mental health. *Behaviour Research and Therapy*, *123*, 103506.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (pp. 4171–4186). Minneapolis, Minnesota. Association for Computational Linguistics.

Dudycha, L. W., & Naylor, J. C. (1966). Characteristics of the human inference process in complex choice behavior situations. *Organizational Behavior and Human Performance, 1*, 110-128.

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., ... & Schwartz, H. A. (2018). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences, 115*(44), 11203-11208.

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Thomson Brooks/Cole Publishing Co.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Ganesan, A. V., Matero, M., Ravula, A. R., Vu, H., & Schwartz, H. A. (2021, June). Empirical evaluation of pre-trained transformers for human-level NLP: The role of sample size and dimensionality. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting* (p. 4515). NIH Public Access.

Garb, H. N., & Wood, J. M. (2019). Methodological advances in statistical prediction. *Psychological Assessment, 31*(12), 1456–1466.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin, 73*(6), 422–432.

Goldberg, L. R. (1976). Man versus model of man: Just how conflicting is the evidence? *Organizational Behavior and Human Performance, 16*, 13-22.

Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H. C., & Jeste, D. V. (2019). Artificial intelligence for mental health and mental illnesses: an overview. *Current psychiatry reports, 21*(11), 1-18.

Grebstein, L. C. (1963). Relative accuracy of actuarial prediction, experienced clinicians, and graduate students in a clinical judgment task. *Journal of Consulting Psychology, 27*, 127-132.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, public policy, and law, 2*(2), 293.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment, 12*(1), 19–30.

Gustafson, D. H., Greist, J. H., Stauss, F. F., Erdman, H., & Laughren, T. (1977). A probabilistic system for identifying suicide attemptors. *Computers and Biomedical Research*, *10*(2), 83-89.

Habli, I., Lawton, T., & Porter, Z. (2020). Artificial intelligence in health care: accountability and safety. *Bulletin of the World Health Organization*, *98*(4), 251.

Heaton, R. K., Grant, I., Anthony, W. Z., & Lehman, R. A. (1981). A comparison of clinical and automated interpretation of the Halstead-Reitan battery. *Journal of Clinical Neuropsychology, 3*, 121-141.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science, 349*(6245), 261-266.

Holt, R. R. (1958). Clinical and statistical prediction: A reformulation and some new data. *Journal of Abnormal and Social Psychology, 56*, 1-12.

Holt, R. R. (1970). Yet another look at clinical and statistical prediction: Or, is clinical psychology worthwhile? *American Psychologist, 25*, 337-349.

Kelly, E. L., & Fiske, D. W. (1950). The prediction of success in the VA training program in clinical psychology. *American Psychologist, 5*, 395-406.

Kjell, O. N. E., Daukantaitė, D., & Sikström, S. (2021). Computational language assessments of harmony in life—not satisfaction with life or rating scales—correlate with cooperative behaviors. *Frontiers in psychology, 12*, 601679.

Kjell, O., Giorgi, S., & Schwartz, H. A. (2023). The text -package: An R-package for analyzing and visualizing human language using natural language processing and transformers. *Psychological Methods*. Advance online publication.

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.

Kjell, O. N. E., Kjell, K., & Schwartz, H. A. (2023, January 4). AI-based Large Language Models are Ready to Transform Psychological Health Assessment. *https://doi.org/10.31234/osf.io/yfd8g*

Kjell, K., Johnsson, P., & Sikström, S. (2021). Freely generated word responses analyzed with artificial intelligence predict self-reported symptoms of depression, anxiety, and worry. *Frontiers in Psychology, 12*, 602581.

Kjell, O. N. E., Sikström, S., Kjell, K., & Schwartz, H. A. (2022). Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific reports, 12*(1), 1-9.

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of chiropractic medicine, 15*(2), 155–163.

Lee, I. A., & Preacher, K. J. (2013). Calculation for the test of the difference between two dependent correlations with one variable in common [Software]. Retrieved from http://www.quantpsy.org/corrtest/corrtest2.htm

Lee, E. E., Torous, J., De Choudhury, M., Depp, C. A., Graham, S. A., Kim, H. C., Paulus, M. P., Krystal, J. H., & Jeste, D. V. (2021). Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom. *Biological psychiatry: Cognitive neuroscience and neuroimaging, 6*(9), 856–864.

Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. *The Alan Turing Institute*.

Libby, R. (1976a). Man versus model of man: Some conflicting evidence. *Organizational Behavior and Human Performance, 16*, 1-12.

Libby, R. (1976b). Man versus model of man: The need for a nonlinear model. *Organizational Behavior and Human Performance, 16*, 23-26.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692.*

Mann, R. D. (1956). A critique of PE Meehl's Clinical versus statistical prediction. *Behavioral Science, 1*(3), 224-230.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276-282.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press, Washington, DC.

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment, 50*, 370-375.

Miller, D. J., Spengler, E. S., & Spengler, P. M. (2015). A meta-analysis of confidence and judgment accuracy in clinical decision making. *Journal of Counseling Psychology, 62*, 553–567.

Nangia, N., & Bowman, S. R. (2019). Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. *arXiv preprint arXiv:1905.10425*.

Omron, R., Kotwal, S., Garibaldi, B. T., & Newman-Toker, D. E. (2018). The diagnostic performance feedback “calibration gap”: why clinical experience alone is not enough to prevent serious diagnostic errors. *AEM education and training*, *2*(4), 339-342.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning* (No. 47). University of Illinois press.

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*(1), 547-577.

Preoţiuc-Pietro, D., Schwartz, H. A., Park, G., Eichstaedt, J., Kern, M., Ungar, L., & Shulman, E. (2016). Modeling valence and arousal in facebook posts. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 9-15).

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology, 39*(6), 1161.

Saposnik, G., Redelmeier, D., Ruff, C. C., & Tobler, P. N. (2016). Cognitive biases associated with medical decisions: a systematic review. *BMC medical informatics and decision making, 16*(1), 1-14.

Schiff, G. D. (2008). Minimizing diagnostic error: the importance of follow-up and feedback. *The American journal of medicine*, *121*(5), S38-S42.

Schwartz, H. A., Eichstaedt, J., Blanco, E., Dziurzynski, L., Kern, M., Ramones, S., ... & Ungar, L. (2013, June). Choosing the right words: Characterizing and reducing error of the word count approach. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity* (pp. 296-305).

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one, 8*(9), e73791.

Shrout, P. E., & Fleiss, J. L. (1979), Intraclass correlation: uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Sun, J., Schwartz, H. A., Son, Y., Kern, M. L., & Vazire, S. (2020). The language of well-being: Tracking fluctuations in emotion experience through everyday speech. Journal of Personality and *Social Psychology, 118*(2), 364.

Tanana, M. J., Soma, C. S., Kuo, P. B., Bertagnolli, N. M., Dembe, A., Pace, B. T., ... & Imel, Z. E. (2021). How do you feel? Using natural language processing to automatically rate emotion in psychotherapy. *Behavior research methods*, 1-14.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24-54.

Topol E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine, 25*(1), 44–56.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*, 5998–6008.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods, 45*(4), 1191-1207.

Wiggins, N., & Kohen, E. S. (1971). Man vs. model of man revisited: The forecasting of graduate school success. *Journal of Personality and Social Psychology, 19*, 100-106.

Wright, W. F. (1979). Properties of judgment models in a financial setting. *Organizational Behavior and Human Performance, 23*, 73-85.

Appendix A

Instructions for condition 1

You are invited to take part in a study that is investigating emotion, and concerns how people respond to different types of words. You will use a scale to rate how you felt while reading each word. The scale ranges from 1 (unhappy) to 9 (happy). At one extreme of this scale (1), you feel completely unhappy, and at the other end of the scale (9), you feel completely happy. If you feel completely neutral, neither happy nor sad, select the middle of the scale (5). Next to this, you are asked to rate your confidence in the assessment of each word with a percentage of 50, 60, 70, 80, 90 or 100. A score of 50% means that you do not feel confident at all about your assessment and 100% means that you feel extremely confident about your assessment. Please perform the ratings in an environment without any distractions, work at a rapid pace and don't spend too much time thinking about each word.

Instructions for condition 2

You are invited to take part in a study that is investigating the valence or sentiment of Facebook posts. Valence or sentiment represents the polarity of the affective content in a post, ranging from very negative (1) to very positive (9). You are asked to rate each post on a nine-point scale, ranging from 1 (very negative) to 5 (neutral/objective) to 9 (very positive). Please perform the ratings in an environment without any distractions, work at a rapid pace and don't spend too much time thinking about each post.

Appendix B

Descriptive Statistics for the Assessments of Texts (condition 2)

Assessment Characteristics	Person A		Person B		Person C	
	Mean (SD)	Median (Min - Max)	Mean (SD)	Median (Min - Max)	Mean (SD)	Median (Min - Max)
Valence Person	5.33 (1.81)	5 (1 - 9)	5.30 (1.09)	5 (2 - 9)	5.26 (1.53)	5 (1 - 9)
Valence Person's Model	5.30 (1.43)	5.16 (0.97 - 9.43)	5.27 (0.74)	5.21 (3.45 - 7.36)	5.26 (1.03)	5.21 (2.09 - 8.50)
Valence Reference Standard (*M*) [a]	5.32 (1.41)	5.40 (1.60 - 8.60)	5.33 (1.55)	5.40 (1.40 - 8.80)	5.34 (1.47)	5.40 (1.40 - 8.60)
Variability among assessors (*SD*) [b]	0.72 (0.40)	0.71 (0.00 - 2.12)	0.71 (0.41)	0.63 (0.00 - 2.39)	0.71 (0.39)	0.71 (0.00 - 2.39)

Note. N = 500. The mean (SD) and median (min - max) of the assessments of the three assessors included in the reference standards are 5.28 (1.82) and 5 (1 - 9) for Person D; 5.52 (1.58) and 6 (1-9) for person E; and 5.26 (1.79) and 5 (1 - 9) for Person F.

The inter-rater reliabilities in the reference standards in the assessment of texts were moderate to good (*ICC* = .73, *95% CI* = .70-.76 in the reference standard for Person A; *ICC* = .77, *95% CI* = .74-.80 in the reference standard for Person B; *ICC* = .75, *95% CI* = .73-.78 in the reference standard for Person C). The inter-rater reliabilities in the reference standards for texts with 50 words or more were poor to moderate (*ICC* = .49, *95% CI* = .30-.70 in the reference standard for Person A; *ICC* = .53, *95% CI* = .34-.72 in the reference standard for Person B; *ICC* = .59, *95% CI* = .40-.77 in the reference standard for Person C).

[a] The mean assessment excluding the assessment of the person who's assessments are compared to the reference standard (i.e., the mean of five assessments)

[b] The variability among the assessments (SD) included in the reference standard.