LUNDS UNIVERSITET

Institutionen för psykologi
Kandidatuppsats

Inledande utveckling och validering av språkbaserade bedömningar för meningsfull förändring

Initial Development and Validation of Language-Based Assessments for Meaningful Change

Ulrika Söderström

Kandidatuppsats HT 23

Handledare: Oscar Kjell
Examinator: Henrik Levinsson

Abstract

Meaningful change has been discussed in multiple studies, with the recurring question of how it could be conceptualized and assessed to identify what determines meaningful change and where it occurs. Previous studies have conducted statistical analyses based on traditional rating scales (i.e., the PHQ-9) to assess meaningful change. There is no evidence to be found of previous studies attempting to assess meaningful change through language-based assessments. This study intended to examine whether language-based assessments could be utilized in assessing meaningful change, and if so, to what extent. This study has utilized scores from human-rated meaningful change assessments of natural language responses (NLR) and self-reported scores from the open-source Patient Health Questionnaire-9 (PHQ-9). The study conducted analyses in R-studio based on the *text*-package and included the large-language model RoBERTa for word embedding, correlation testing for examining reliability and validity, and ridge regression to train the model. The analyses showed results of inter-rater reliability in human-rated assessments (r = .64, p <.001, N = 100), correlation between human-rated assessments and PHQ-9 difference scores (r = .36, p <.001, N = 298), the strongest trained model (r = .39, p <.001, N = 298), and correlation between language-based assessment and PHQ-9 difference scores (r = .29, p <.001, N = 298). These findings suggest that language-based assessments can be further developed to assess meaningful change, and preferably by including human-rated assessment.

Keywords: meaningful change, depression, Large Language Models, AI

Sammanfattning

Meningsfull förändring har diskuterats i flera studier, med den återkommande frågan om hur den skulle kunna konceptualiseras och bedömas för att identifiera vad som utgör meningsfull förändring och var den uppstår. Tidigare studier har genomfört statistiska analyser baserade på traditionella betygsskalor (d.v.s. PHQ-9) för att bedöma meningsfull förändring. Det finns inga bevis för att tidigare studier försökt bedöma meningsfull förändring genom språkbaserade bedömningar. Denna studie syftade till att undersöka om språkbaserade bedömningar kan användas för att bedöma meningsfull förändring, och i så fall i vilken utsträckning. Denna studie har använt poäng från mänskligt värderade meningsfulla förändringsbedömningar av natural language responses (NLR) och självrapporterade poäng från Patient Health Questionnaire-9 (PHQ-9). Studien genomförde analyser i R-studio utifrån *text*-paketet och inkluderade large language modellen RoBERTa för word embedding, korrelationstestning för att undersöka reliabilitet och validitet samt ridge regression för att träna modellen. Analyserna visade resultat av reliabilitet mellan de mänskliga bedömarna (r = .64, p <.001, N = 100), korrelation mellan mänskliga bedömningar och skillnadspoäng från PHQ-9 (r = .36, p <.001, N = 298), den starkast tränade modellen (r = .39, p <.001, N = 298), och korrelation mellan språkbaserade bedömningar och skillnadspoäng från PHQ-9 (r = .29, p <.001, N = 298). Dessa fynd tyder på att språkbaserade bedömningar kan vidareutvecklas för att bedöma meningsfull förändring, och helst genom att inkludera mänskliga bedömningar.
Nyckelord: meaningful change, depression, Large Language Models, AI

Acknowledgements

I would like to start to thank all participants for their valuable contribution. A special thank you to my supervisor Oscar Kjell for making this study possible, your help has been invaluable. I would also like to thank my family and friends for your patience with letting this thesis having my full attention during this period. Lastly, I would like to thank Kevin and Veerle from the Ablemind team for your tremendous help.

Initial Development and Validation of Language-Based Assessments for Meaningful Change

The field of clinical psychology has long been dedicated to understanding human behaviour and promoting positive change in individuals (Schmidt & Power, 2005). The concept of Meaningful Change encompasses changes in thoughts, emotions, and behaviours that lead to improvements in mental health that is meaningful for the individual (Byrom et al., 2020). Understanding the factors that contribute to meaningful change is crucial for advancing our understanding of psychological processes and developing effective interventions (Byrom et al., 2020; Jacobson et al., 1984; Jacobson & Truax, 1991). This study concerns how to assess what constitutes meaningful change, which means the approach to identify the change between two time points.

The condition this study focuses on is depression. The assessment of depression today is mostly based on traditional rating scales which indicate the severity of a patient's condition through general thresholds (Kroenke et al., 2001). Meaningful change in this context is to measure the change between the scores from the same patient at two time points (Byrom et al., 2020). These scores can then be analysed through various metrics to identify the change, such as effect size or statistical significance. The methods for identifying effect sizes and statistical significance can be considered valuable tools for identifying change, but there are limitations regarding their ability to assess *meaningful* change (Eisen et al., 2007). Eisen et al. (2007) highlighted that "with reasonably large samples, it is possible for small differences that may not be *clinically meaningful* to reach statistical significance." (p. 273, italics added). Further, "while the importance of statistical significance in demonstrating the effects of an intervention is unquestioned, it is also important to recognize that effect sizes detected through statistical tests may be of insufficient magnitude to be considered relevant to the patient." (Byrom et al., 2020, p. 3). These concerns indicate the need for further developing ways of assessing *meaningful* change.

In recent years, psychology research has witnessed significant advancements taking place in how psychological constructs are measured using natural language analyses. This can be seen in the form of large language models, which are powerful artificial intelligence (AI) systems that are capable of processing and generating natural language (Cheng et al., 2023; Otsuka et al., 2023). Language-based assessments have shown potential in assessing

psychological constructs, and a study by Kjell et al. (2022) investigate the use of AI-based transformers to analyse natural language and predict subjective well-being measures. The study demonstrates that these AI models can achieve remarkably high levels of accuracy, approaching the theoretical upper limits, in predicting traditional subjective well-being measures. In the context of reliability, the theoretical upper limit represents the maximum level of performance or functioning that can be achieved under ideal conditions. It can serve as a reference point for evaluating the actual reliability of a system or process (Kjell et al., 2022). The findings suggest that AI-powered analysis of natural language has the potential to provide valuable insights into individuals' well-being, offering a promising avenue for future research and applications in this field. These unexplored potentials are the foundation of the novel idea of this study, and thanks to previous work, there is now an opportunity for a step in how to assess meaningful change in natural language (Kjell et al., 2022).

This study will employ AI techniques to analyse open-ended responses in order to assess meaningful change. The primary objective is to investigate the potential of utilizing large language models to construct a model capable of assessing meaningful change. Large language models are trained on extensive text data, enabling them to acquire intricate language patterns and relationships. Consequently, they can support psychologists in various tasks such as sentiment analysis, language translation, text summarization, and even therapeutic interventions (Binz & Schulz, 2023; Cheng et al., 2023; Kjell et al., 2022). The present study focuses on the development and validation of language-based assessments to evaluate meaningful change in psychological constructs, specifically targeting depression. This involves analysing changes in descriptions of mental health over time through human assessment, with the aim of identifying the occurrence and magnitude of meaningful change. Subsequently, the study aims to explore the possibility of developing a model based on individuals' own descriptions of their mental health. This exploration will be conducted through analyses utilizing large language models, AI techniques, and human-assessed scores, with the goal of predicting meaningful change.

As the study aims to develop a novel approach to assess meaningful change within an individual, it can be seen as a valuable contribution to the search of a more accurate method for assessing treatment effectiveness and determining the most suitable treatment approach for each

patient. To be able to determine what meaningful change is, an initial exploration of the definition of change in the context of clinical psychology will be presented.

Change in psychological assessments

"Change" refers to any observable difference in a particular variable or outcome measure. It can be seen as a change in scores of a psychological assessment or as a change in behaviour or symptom severity (Byrom et al., 2020; Guyatt et al., 1987).

The assessment of change can be seen as a way to measure the size and significance of differences between groups as well as within individuals. By comparing, for example, pre- and post-intervention scores, researchers can determine the magnitude and statistical significance of change (Jacobson et al., 1984). These assessments aim to capture changes in various psychological domains, such as cognition, emotion, behaviour, and overall mental health (Jacobson et al., 1984; Jacobson & Truax, 1991). To effectively measure change, assessments often employ quantitative measures, such as standardized questionnaires or rating scales, which provide numerical data that can be analysed statistically.

There are different approaches to assess change, and some of the most commonly used methods today are the Reliable Change Index (RCI), Standard Error of Measurement (SEM), Cohen's *d*, and Smallest Effect Size of Interest (SESOI; Cohen, 1988; Gruijters & Peters, 2022; Jacobson & Truax, 1991; McHorney & Tarlov, 1995). These developed methods are validated by empirical evidence (see Cohen, 1988; Gruijters & Peters, 2022; Jacobson & Truax, 1991; and McHorney & Tarlov, 1995).

Statistical significance

The Reliable Change Index (RCI). The Reliable Change Index (RCI) is a statistical measure to determine whether an individual's change in a particular variable, such as symptoms or behaviours, is statistically significant (Jacobson & Truax, 1991). It helps to assess whether the observed change is beyond what would be expected due to measurement error or random fluctuations. The RCI takes into account both the magnitude of change and the variability of the measurement instrument used (Jacobson et al., 1984; Jacobson & Truax, 1991). It compares an individual's pre- and post-scores on a specific measure and calculates the extent to which the change exceeds what would be expected by chance. This calculation considers factors such as the reliability of the measurement instrument and the standard deviation of the scores (Jacobson et al., 1984; Jacobson & Truax, 1991).

The Standard Error of Measurement (SEM). The Standard Error of Measurement (SEM) is a statistical concept to estimate the amount of error inherent in a test or measurement. It provides an indication of the precision or reliability of a test score by estimating the extent to which an individual's true score may vary from their observed score (McHorney & Tarlov, 1995). The SEM helps us understand the margin of error associated with a test score. It quantifies the amount of variability we can expect in an individual's scores if they were to take the same test multiple times under similar conditions (McHorney & Tarlov, 1995). The SEM is valuable in evaluating the effectiveness of interventions or educational programs by assessing whether observed changes in scores exceed the measurement error, compared to RCI which focuses on statistical significance and reliability (Jacobson & Truax, 1991; McHorney & Tarlov, 1995). It is a crucial statistical concept in psychometrics that aids in the interpretation and evaluation of test scores (McHorney & Tarlov, 1995).

Effect sizes

Cohen's *d*. Cohen's *d* is a statistical measure that quantifies the effect size (ES) of the difference between two test scores. It is commonly used in research to determine the magnitude of the difference between means and to assess the size of the findings. Interpreting Cohen's *d* values can vary depending on the field of study and the specific context. However, a commonly used guideline is that a Cohen's *d* value of 0.2 is considered a small effect size, 0.5 is considered a medium effect size, and 0.8 or above is considered a large effect size (Cohen, 1988; Wolters Kluwer Health, 2000). In the context of meaningful change, these thresholds are referred as common but not optimal to determine the value of meaningful change based on the effect size (Gruijters & Peters, 2022). This due to the subjective aspects of how one conceptualizes meaningful change.

The Smallest Effect Size of Interest (SESOI). The Smallest Effect Size of Interest (SESOI) refers to the minimum significant difference or effect that researchers or practitioners are interested in detecting. It is a threshold that helps determine whether an observed effect or difference is considered meaningful or not. SESOI is typically defined based on the context and goals of the study, and it can vary across different fields and research areas. Its purpose is to distinguish between statistically significant findings and those that have practical significance or real-world relevance (Gruijters & Peters, 2022). One big difference between this method and

the RCI, SEM, and ES is that SESOI focuses on trying to find a threshold of how small a change can be to be seen as interesting (Byrom et al., 2020; Gruijters & Peters, 2022).

Evaluation of the methods

Eisen et al. (2007) took a closer look at the reliable change index, standard error measurement, and effect size to examine if they could determine the approach that yields a clinically significant change estimate that is most consistent with other change measurements. It was shown that "both the SEM and ES methods identified a higher proportion of individuals as meaningfully improved than did the RCI method." (Eisen et al., 2007, p. 286). The result of their study indicates that SEM was the most efficient way to assess change and move closer to the assessment of meaningful change. This conclusion was based on that SEM had the highest conformity with clinically meaningful improvement and decrease in the mental health scores of the BASIS-24 rating scale (Eisen et al., 2007).

Byrom et al. (2020) pointed out limitations regarding these distribution-based methods: "Despite their inherent simplicity, distribution-based methods, however, fail to associate statistical changes with whether a truly meaningful change has occurred." (p. 4). This is due to their lack of ability to determine estimates based on small sample sizes or baseline data with large variability.

Meaningful change in psychological assessments

"Meaningful change" goes beyond statistically significant differences and refers to changes that are considered relevant to the individual's mental health or functioning. Determining meaningful change can be challenging, especially due to the various names used for the same phenomenon. Clinically Important Difference (CID), Minimal Important Difference (MID), and Minimally Clinically Important Difference (MCID) are three different ways to address the minimum change that is seen as clinically relevant (Byrom et al., 2020). In other words, what indicates the smallest change that could be of interest.

Meaningful change takes into account the practical implications and real-world impact of the observed changes (Byrom et al., 2020). It considers whether the change is substantial enough to make a difference in a person's life, treatment outcomes, or overall psychological health (Byrom et al., 2020). Byrom et al. (2020) writes "Meaningful change can be considered to represent the smallest difference in an endpoint measure that would be perceived by patients as beneficial." (p. 3). One way to describe the difference between the assessment of *change* and

the assessment of *meaningful change* is by a method's ability to identify if a meaningful change can be seen (Byrom et al., 2020). Different individuals may have varying interpretations of what is considered meaningful in their lives (Byrom et al., 2020; Jacobson et al., 1984; Jacobson & Truax, 1991).

Byrom et al. (2020) illuminates different methods for measuring meaningful change and divides them into three main domains: 1) Consensus-based methods, 2) Distribution-based methods, and 3) Anchor-based methods.

Consensus-based methods can be described as when researchers, together with the parties involved, make an agreement regarding the expected result, which then "define a threshold for clinically relevant change in the specific patient population to be studied" (Byrom, et al., 2020, p. 3). This study can be categorized in this domain due to its attempt of contributing to define a threshold to determine meaningful change. Distribution-based methods rely on analysing the distribution of the recorded outcome measure to determine the extent of change that is unlikely to occur randomly. It is common to use multiple distributional methods to establish a consensus or range for the minimally clinically important difference value (Byrom et al., 2020), and the earlier presented measurements (ES, RCI, SEM, and SESOI) can be placed in this domain. It is important to highlight that these measurements are not measures of meaningful change in themselves but can be attached as thresholds for indicating meaningful change. Anchor-based methods in measuring meaningful change refer to a statistical approach that utilizes external reference points, known as anchors, to assess the magnitude of change in a specific variable or construct. These anchors can be objective measures, such as clinical assessments or established rating scales, that are considered to be valid indicators of the concept being measured. By comparing the scores obtained before and after an intervention or over a certain period of time, anchor-based methods help determine the extent of meaningful change by examining the relationship between the change scores and the anchor scores. This approach provides a framework for interpreting and quantifying the significance of observed changes in the given context (Byrom et al., 2020).

It is important to note that assessments for meaningful change should be valid, reliable, and sensitive to individual differences. Validity ensures that the assessment measures what it intends to measure, while reliability ensures consistency and stability of the measurement over time. Sensitivity to individual differences acknowledges that meaningful change can manifest

differently for each person, and assessments should be able to capture these individual variations (Byrom et al., 2020).

Even though meaningful change initially refers to the experienced change within an individual, it can also be assessed in relation to, for example, cost-benefit aspects (Gruijters & Peters, 2022). Gruijters and Peters (2022) discuss the effect size of what is meaningful in relation to the "consideration of cost-benefit", which refers to that what is meaningful is based on the effect size of the outcome in relation to the invested costs of the study or treatment.

Limitations with current ways of conceptualizing meaningful change

Even though the current ways of conceptualizing meaningful change in psychology have their strengths, there are also some limitations to consider. One limitation is the subjective nature of defining what constitutes meaningful change. So, it is important to consider the contextual factors that shape an individual's experiences and the potential limitations of solely focusing on significance tests and effect sizes. Significance tests and effect sizes do not take these factors into account in relation to the potential that meaningful change may vary across contexts (Blampied, 2022; Jacobson & Truax, 1991; McAleavey, 2021).

Furthermore, the assessment of meaningful change can be challenging. Objective and standardized measures may not fully capture the complexity and nuances of personal transformation. Overall, while current conceptualizations of meaningful change provide valuable insights, ongoing research and exploration are necessary to address these limitations of context and further enhance our understanding of meaningful change as this important psychological phenomenon (Byrom et al., 2020; McAleavey, 2021).

Even though the conclusions of these studies showed the RCI as valuable to a certain extent, the limitation regarding where the meaningfulness can be identified in the change still stands (Blampied, 2022; McAleavey, 2021). The need of further assessment to determine the meaningfulness is discussed by Blampied (2022) saying that "it goes without saying, that no decision to amend, change or terminate therapy or to classify a client in some way (e.g., as fit for work) should be made on the basis of RCI alone" (p. 14). Which can be interpreted as the need of developing a method or tool to complement the statistics.

Here is where this study's conceptualization of meaningful change enters with the idea to base the assessments on language instead of rating scores. The aim is to determine the placement of meaningfulness through human assessment of the natural language, this to be able

to potentially identify meaningful change that would have been overseen in the assessment tools and methods available today. But before we are moving on to discuss this language-based assessment it is necessary to look into the current ways of collecting the data that these above-mentioned assessment methods are based on.

Rating scales for assessment of mental health

The data that change and meaningful change have been assessed on is solely the scores of different rating scales (Blampied, 2022; Bost et al., 2008; Byrom et al., 2020; Eisen et al., 2007; Jacobson & Truax, 1991; McAleavey, 2021; McHorney & Tarlov, 1995). Rating scales are commonly used tools for measuring and quantifying subjective experiences, behaviours, or attitudes (Wolters Kluwer Health, 2000). These scales provide a structured way to assess various psychological constructs, such as personality traits, mental health symptoms, or well-being. They typically consist of a series of statements or items that individuals rate based on their agreement, frequency, intensity, or other relevant dimensions (Wolters Kluwer Health, 2000).

The rating scales can be self-report measures completed by individuals themselves or observer-rated measures completed by trained professionals. They offer a standardized and reliable method for collecting data, allowing for comparisons and statistical analysis as the before mentioned methods (Uher, 2023). By using rating scales, psychologists can gain valuable insights into individuals' thoughts, feelings, and behaviours, aiding in research, diagnosis, and treatment planning (Uher, 2023).

Rating scales have important implications for healthcare providers in terms of facilitating early detection, monitoring treatment efficacy, and enhancing patient outcomes. When used as part of routine screening, various rating scales can help clinicians identify patients who may benefit from further evaluation or intervention (Uher, 2023). By promptly identifying depression symptoms, healthcare providers can initiate appropriate treatments, such as psychotherapy or pharmacotherapy, to alleviate symptoms and prevent the worsening of the condition (Wolters Kluwer Health, 2000).

The use of rating scales in relation to change and meaningful change can appear valuable for the assessor, it can even seem like the ratings are valid in the context of assessing change. But the limitations within this assessment can be identified in the result of a statistical analysis based on the ratings from two time points from the same individual (Uher, 2023). To put this

into a more comprehendible context have we made an example based on one of the many rating scales, the Patient Health Questionnaire-9 (PHQ-9). The example is displayed in Table1.

Table 1

Example of how the scores of a PHQ-9 can look

	PHQ-9 questions	PHQ-9 T1	PHQ-9 T2
1.	Little interest or pleasure in doing things	2	2
2.	Feeling down, depressed, or hopeless	1	1
3.	Trouble falling or staying asleep, or sleeping too much	**3**	**0**
4.	Feeling tired or having little energy	2	2
5.	Poor appetite or overeating	0	0
6.	Feeling bad about yourself – or that you are a failure or have let yourself or your family down	1	1
7.	Trouble concentrating on things, such as reading the newspaper or watching television	1	1
8.	Moving or speaking so slowly that other people could have noticed? Or the opposite – being to fidgety or restless that you have been moving around a lot more than usual	1	1
9.	Thoughts that you would be better off dead or hurting yourself in some way	**0**	**3**
	Total Score:	11	11

PHQ-9 T1 = Time point one; PHQ-9 T2 = Time point two; Total Score = The hypothetical score from all nine questions; Interpretation = 0 – not at all, 1 – several days, 2 – more than half of the days, 3 – nearly every day; Total score = 0–4 none, 5–9 mild depression, 10–14 moderate depression, 15–19 moderate severe depression, and 20–27 severe depression.

According to the total score of 11 for both T1 and T2, there has not been any change between time point one and time point two, and the severity of the depression state indicates a

moderate depression. But if you look more closely at the questions you can see that the patient at T1 was mostly tired and had troubles regarding sleeping, while at T2 the patient is indicating to be suicidal. By the change in these specific areas, you can see that something most likely has happened. This is as mentioned previously is a way for typical rating scales in a statistical context to potentially overlook change that can be of meaning for the individual, and importance for a psychologist to provide with the correct treatment. Uher (2023) highlights the lack of attention to language and its value in the traditional rating scales (Uher, 2023).

To enlighten this study's aim and put it into context have we made an example of the before mentioned idea of assessing language instead of rating scores. In Table 2 you can see a brief presentation of how it can look when using natural language responses for the assessment.

Table 2

Example of a Natural Language Response

Mental health description question	NLR T1	NLR T2
Please describe how you have been over the last two weeks	I have had sleep problems, feeling low. But I feel good about my job and my colleagues which have given me energy to get through the day even though I feel very tired.	My sleep is good, but I have problems at my job. I am so tired of a new colleague's low work ethics which affects me every day. This is just draining me on energy, and I feel like that might be the reason for me being able to sleep through the night.

NLR T1 = Natural Language Response at time point one; NLR T2 = Natural Language Response at time point two.

When looking at these natural language responses, you can see that a change has occurred. The individual uses the same or similar words to explain the condition, but when you look at the individual responses at T1 "I have sleep problems, feeling low. But I feel good about my job" and at T2 "My sleep is good but now I have problems at my job", you can see that the words explain two different conditions with two different meanings for the individual. This is an example of what the human assessed rating scores can be based on to build the language-

based assessment model. Since the purpose of this study is to explore ways to assess meaningful change, we are now moving away from these numeric based ways of collecting information about the patient's mental health and moving over to introducing the large-language models.

Language-based assessment using large language model

These large language models, fuelled by the power of deep learning and vast amounts of data, have developed the way machines understand and generate human language (Kjell et al., 2023a). With their ability to process and generate text in a more human-like manner, large language models have opened up new possibilities for various applications, ranging from chatbots and virtual assistants to content generation and language translation (Cheng et al., 2023; Otsuka et al., 2023). A fundamental building block of large language models is the Transformer architecture (Vaswani et al., 2017). This architecture enables the models to capture contextual relationships and language patterns effectively (Vaswani et al., 2017).

Large language models undergo a process of pre-training. During pre-training, the models learn general language patterns and relationships by processing vast amounts of unlabelled text data through unsupervised learning (Cheng et al., 2023; Devlin et al., 2019; Otsuka et al., 2023). In our case, for assessing meaningful change.

Multiple studies have explored the possibilities of using natural language in clinical assessments and the measurement of mental health (Kjell et al., 2022; Kjell et al., 2023a; Kjell et al., 2023b). Studies have shown that language assessments using large language models might be more valid than rating scales in assessing psychological constructs (Kjell et al., 2019; Kjell et al., 2023b; Kjell et al., 2022; Sikström et al., 2023). These studies support the potential of developing a language-based assessment tool to predict meaningful change is worth exploring. The current knowledge is that language-based assessments have not yet been developed for meaningful change, and the assessment tools today are typically based on the data from rating scales.

Assessments of meaningful change play a crucial role in measuring and evaluating the changes that individuals experience, and these indications can potentially help psychologists and clinicians provide the most suitable treatment for future patients (Byrom et al., 2020). So, we aim to create a model that could potentially change the way meaningful change can be assessed through language-based assessment.

Hypotheses

The research question includes examining to what extent language-based assessments can be used to improve the way meaningful change is assessed. Hypothesis 1 is for establishing the validity and reliability of the human-rated meaningful change assessments. Hypothesis 2 is about developing the language-based meaningful change models, and hypothesis 3 is about validating the language-based assessments.

Hypothesis 1

Human assessed level of meaningful change in depression based on descriptions of mental health at two different time points **a)** show strong inter-rater reliability, and **b)** significantly correlate with the difference scores of self-reported depression as measured by a traditional rating scale (i.e., the PHQ-9).

Hypothesis 2

Individuals' mental health descriptions at two different time points predict **a)** human assessed level of meaningful change in depression and **b)** the difference scores of self-reported depression as measured by a traditional rating scale (i.e., the PHQ-9).

Hypothesis 3

Language-based assessments of meaningful change significantly correlates positively with the difference scores of a self-reported depression scale (PHQ-9) for predicting meaningful change.

Method

Participants

The mental health descriptions were collected in a previous study regarding mental health (Kjell et al., in progress). The participants were recruited from a participant recruiting platform called Prolific (https://www.prolific.com/) and received a reward of 7.5 pounds/hour for their participation. The sample contains 298 English speaking participants from the UK, with successfully completed answers. There are 192 females, 105 males and 1 gender unknown participants. The age mean is 47.4 years (SD = 18.1, 95% CI [18–84], N = 298).

Raters

The mental health descriptions were assessed by two students from Lund University with the age of 29 and 28 years. Student number 1 (S1) is a master's student in psychology.

Student number 2 (S2) is a Bachelor student in psychology and who also is the author of this thesis.

Instruments

Mental health description

The mental health question: "How is your mental health?". With the instructions "Please describe how you have been over the last two weeks. You can, for example, write about your emotions, thoughts, behaviours, and/or symptoms related to your health", "Write at least one paragraph", and "Please answer with at least 20 words (and no more than 1000)". The question was inspired from Kjell et al. (2019).

PHQ-9

The rating scores used Patient Health Questionnaire-9 (PHQ-9). One of the primary objectives of the PHQ-9 are to identify individuals who may be experiencing depression or related mood disorders (Kroenke et al., 2001). By incorporating a comprehensive set of questions covering the nine diagnostic criteria for major depressive disorder outlined in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), the PHQ-9 provides a systematic approach to screening for depressive symptoms. The nine questions have then been self-rated by the participants on a scale from 0 (not at all) to 3 (nearly every day) (Kroenke et al., 2001). The scores used in this study had a Cronbach's alpha of .87, which indicates good internal consistency.

Meaningful change assessment questions

To assess the natural language descriptions, the meaningful change questionnaire was developed by S1 of the study. These questions give us the Meaningful change assessment scores. It is developed with the purpose of assessing meaningful change and valence between T1 and T2 based on the participants' descriptions in written text. The given instructions for the approach of the assessment are:

"In this study, people have been asked to describe their general mental health, first at T1 and then again approximately two months later at T2. The individuals have not been influenced in any way, and no intervention was implemented between the two points of measurement. Please rate whether the change of general mental health (if any) was meaningful for the individual or not. After that, please indicate the valence of the change in general mental health (if any) between T1 and T2. Please remember that you are rating any change for that

individual alone; think about the valence and meaningfulness of the change from T1 to T2 of that individual and try not to compare them to other individuals in the dataset.

Ratings for Meaningfulness: Please rate whether the change in general mental health (if any) was meaningful for the individual or not. Where 0 means that the change was not meaningful at all, and 7 means it was very meaningful.

Ratings for Valence: Please indicate the valence of the change in general mental health (if any) between T1 and T2. Where -5 means that there is a very negative change, 0 means no change, and 5 means that there is a very positive change."

Procedure

The data was collected in an online survey, where participants first were given information about the study, their right to withdraw at any time without giving a reason and provided consent to participate in the study. The participants were asked to describe their mental health with above mentioned open-ended questions, followed by filling out rating scales, including the PHQ-9 at two time points, T1 and T2. The time between T1 and T2 was approximately two months, and the survey took approximately 30 minutes in total.

The mental health descriptions were assessed by two students collaborating to create their own separate thesis and started with randomizing and dividing the participants between S1 and S2, with an individual sample size of approximately 300 participants, with 100 overlapping. This step was important to avoid input bias and to be able to analyse inter-rater reliability. The final analysed sample for S2 contained 298 participants.

The raters then assessed these natural language descriptions trying to identify if it occurred a change and how much that change could mean for that specific individual. These assessments were collected through the above-mentioned meaningful change assessment questions rating scale and created the final dataset for the statistical analyses.

Ethical considerations

Participants

Ethical considerations regarding the participants such as, informed consent, no physical or psychological harm, complete anonymity, and right to withdraw participation without consequences were fulfilled and approved by the Swedish Ethical Review Authority with ID number: Dnr 2021–01820. All data and information about the participants were anonymized

before it was handed over to S1 and S2, and the treatment of the data followed the General Data Protection Regulation (GDPR).

Raters

The aspects that have been considered regarding the raters are consequences to the task of reading and coding the participants' answers, as these can be perceived as offensive and affect the researcher's mental state. This was remedied by ensuring that the raters had conversational support and made sure to take frequent breaks during assessment. The researchers were informed of this and gave consent to carry on with the study under these premises.

Data analysis

The analyses were carried out in R using R-studio (R Core Team, 2023). The R-package used for the analyses was the *text*-package (Kjell et al., 2023a). The study used a significance level of $\alpha = .05$ and the thresholds for the Pearson product-moment correlation coefficient were set as between .1 and .3 (weak), between .3 and .5 (moderate), and greater than .5 (strong) (Cohen, 1988).

Analyses of natural language

The *text*-package is a tool for analysing text. It provides a range of functions and algorithms specifically designed for analysing textual data and it utilizes Natural Language Processing (NLP), Deep Learning, and transformers to enable various text analysis tasks (Kjell et al., 2023a). NLP involves the application of computational techniques to understand and process human language. Deep Learning, a subset of machine learning, utilizes neural networks with multiple layers to extract complex patterns and representations from data. Transformers, a specific type of deep learning model, have developed NLP by effectively capturing contextual relationships between words (Kjell et al., 2023a).

In the *text*-package, NLP techniques are employed to pre-process and clean text data, such as tokenizing. Deep Learning models, including transformers, with their attention mechanisms, excel at capturing long-range dependencies and contextual information, making them particularly effective for tasks requiring understanding of language nuances (Kjell et al., 2023a). Based on these methods, the first step was to create the word embeddings (i.e., numerical representations of language) for mental health descriptions at T1 and T2.

Pre-trained word embeddings

The text data was transformed into word embeddings, which are numerical representations of words. RoBERTa is a large language model that belongs to the family of transformer-based models, it is a pre-trained model that utilizes a large amount of unlabelled text data to learn contextual representations of words and sentences (Liu et al., 2019).

RoBERTa employs a masked language modelling objective, where it predicts missing words in a sentence, and a next sentence prediction objective, where it determines if two sentences are consecutive in a given text. By training on a vast corpus of text, RoBERTa learns to understand the relationships between words and sentences, capturing contextual information effectively (Liu et al., 2019). RoBERTa has demonstrated state-of-the-art performance on several benchmarks and has become a widely used model in the field of natural language processing (Liu et al., 2019).

Matero et al. (2021) examined if RoBERTa could be useful in relation to depression assessment and presented a positive result of its ability for assessing psychological constructs. So, by using RoBERTa instead of another pre-trained language model, we are aiming to achieve a more correct result of the embedding process combined with the meaningful change assessment scores on depression.

Training word embeddings to rating scales for meaningful change

In order to investigate the correlation between the words/texts and the change assessments scores, the word embedding dimensions of the data were utilized as predictors in ridge regression. This regression analysis aimed to predict the change assessment scores. The training process involved tenfold cross-validation, where the training set was repeatedly divided into analysis sets (to create models with varying penalties), assessment sets (to evaluate the different models), and test sets (to apply the best-evaluated model). The penalty range for ridge regression was set from 10^{-16}-10^{16}, and the accuracy of the predictions made by the different models was assessed using Pearson product-moment correlation coefficient (r) between the observed and predicted scores.

Results

Descriptives

The descriptive data in Table 3 shows the mean, standard deviation, and range of the words in the mental health descriptions. The results show that the number of words in the mental health description T1 got a mean of 50.7 (*SD* = 34.3) and T2 got a mean of 39.6 (*SD* = 27.6).

Table 3

Descriptive data of the number of words (Mean, SD and range)

Variables	Mean	*SD*	Min	Max
Mental health description T1	50.7	34.3	19	345
Mental health description T2	39.6	27.6	17	370

N=298; *SD* = standard deviation; Range = Confidence interval of 95%.

The reliability and validity of the human-rated meaningful change

The inter-rater reliability between the two raters showed a significant, strong positive correlation ($r = .64$, $p < .001$, $N = 100$). This result is based on the overlapping assessments rated by S1 and S2. The meaningful change ratings and the difference scores of a traditional rating scale (PHQ-9) showed a significant, moderate positive correlation ($r = .36$, $p < .001$, $N = 298$). The result is based on the assessments rated by S2 and the difference in the self-reported scores on PHQ-9.

Language-based assessments of meaningful change

The Pearson's *r* was strongest when using the concatenated model with meaningful change assessment rating scores ($r = .39$; Table 4). Separately, Pearson's *r* at T2 ($r = .37$) was stronger than The Pearson's *r* at T1 ($r = .26$). The result of the concatenated model showed that it was possible to predict meaningful change by Meaningful change ratings. As can be seen in Table 4, the strongest model (T1 and T2 concatenated) was statistically significant ($p < .001$). These results are based on the assessments rated by S2 ($N = 298$).

Table 4

Statistical data for the six different trained models (Pearson's r and p-value)

	Pearson's *r*	
Word Embeddings	Meaningful change ratings	PHQ-9 difference ratings
T1	.26***	.28***
T2	.37***	.20***
T1 and T2 concatenated	.39***	.32***

Note. $N = 298$; * = $p < .05$; ** = $p < .01$; *** = $p < .001$; T1 = Word embeddings for time point one; T2 = Word embeddings for time point two; T1 and T2 concatenated = Word embeddings for time point one and two concatenated.

External validity of the language-based assessments of meaningful change

The correlation between language-based assessment of meaningful change and the difference scores of self-reported depression prediction showed a significant, weak positive correlation ($r = .29$, $p < .001$, $N = 298$).

Discussion

The overall results of this study have reached a positive outcome regarding answering the research question, by providing evidence suggesting the possibility to develop and validate language-based assessments for meaningful change and to what extent. The correlation of the inter-rater reliability supported hypothesis 1a as it showed a strong statistically significant correlation. This result can be seen as a valuable contribution to show reliability of the meaningful change assessment instructions, and to validate the assessment ratings made by S2, indicating that the ratings have a good foundation. The inter-rater reliability ($r = .64$) can also be seen as a theoretically upper limit of predictive assessment accuracy (Kjell et al., 2022).

Hypothesis 1b, which aimed to establish the external validity of the assessments, was supported by statistically significant results indicating a moderate level of correlation between human-assessed scores of meaningful change and the PHQ-9 difference scores. This provides evidence for the validity of utilizing human-rated assessments in the context of psychological constructs of depression, as they were compared to self-reported scores from the PHQ-9 depression severity measure. The results of the trained models supported hypothesis 2a, with the concatenated language-based model being the strongest. This suggests that the study's aim of using language-based assessments for meaningful change is achievable and holds potential. In relation to using the inter-rater reliability ($r = .64$) as an upper limit, the result of $r = .39$ can

be considered stronger than if based on the general thresholds of Pearson's *r* as presented in the study by Kjell et al. (2022; Cohen, 1988). The results of the meaningful change ratings at T1 and T2 separately compared to them concatenated, indicates that if the model would be trained on larger amount of data, the model could potentially become stronger.

Hypothesis 2b showed a moderate statistically significant correlation. An interesting interpretation of this result is that the model indicated a stronger performance in assessing meaningful change scores rather than PHQ-9 difference scores, which potentially could be explained by the different main points of the assessment examples in Table 1 and Table 2. This can potentially indicate that the human-rated assessment can be seen as a more suitable approach to assess meaningful change. It is important to address that the ratings assessed by S2 (human-rated) correlated with the PHQ-9 difference scores with $r = .36$, whereas the language-based assessment of meaningful change (the trained model) correlated the strongest with $r = .39$. This indicates that the result of the language-based assessment could be stronger if the model would be trained on more data. So, for now, based on these results it is best to use human-raters, then language-based assessment and last PHQ-9 difference scores. The trained model has the potential to assist in identifying changes in a patient's mental state, reducing the risk of overlooking important changes.

The analysis conducted to investigate hypothesis 3 yielded a statistically significant weak correlation. It is important to note that the difference scores used in this analysis are not intended to measure meaningful change directly, but rather serve as a measure of change. Meaningful change should exhibit a (weak to moderate) correlation with another change score, but not a perfect correlation. The presence of this correlation, however, provides some level of external validity evidence for the human-rated assessment scores of meaningful change. Therefore, it is reasonable for the correlation to be significant, albeit not very strong, as the two measures assess different aspects. This correlation ($r = .29$) could also be compared with the correlation in hypothesis 1b ($r = .36$), and just like the indications in the hypothesis 2b discussion, this result could potentially be stronger if it was based on a larger amount of data. As Uher (2023) described regarding the way psychologists can use rating-scales to get valuable insights in the patients' condition, we wish for this tool to be able to do the same. With the purpose of identifying the patient's condition and providing the correct treatment faster. And

based on these results, it indicates to be possible since the language-based assessments indicated at being better at predicting the assessment than a traditional rating scale.

In response to the need for an additional assessment method to complement significance tests, our assessment managed to capture important changes that could potentially be overlooked by scores from a traditional rating scale (Blampied, 2022). Building upon the limitations of self-reported scores in significance tests, this study has identified a potentially importance of having human-rated assessment on the language-based assessment. Which can be seen as an interpretation of Byrom et al.'s (2020) discussion on the limitations of current methods in identifying meaningful change.

The inter-rater reliability results demonstrate a consistent quality in the assessment scores between the assessors. While this method shows promise as an alternative to rating scales, it is in future research crucial to acknowledge its time-consuming nature and consider cost-benefit aspects, as highlighted by Gruijters and Peters (2022). In this regard, as mentioned above, the trained model based on human-rated assessments can serve as a valuable tool to predict assessment outcomes based on patients' natural language responses to open-ended questions. So instead of a human being required to make the assessments of meaningful change, the model can predict what the human assessment would have been. This also eliminates the need for patients to complete extensive rating scales that may not capture their intended message. Compared to rating scales, this model could be seen as adding an additional dimension to the assessment process, bringing us closer to identifying the meaningfulness for the individual.

Envisioned utilization

Since meaningful change aims to achieve a deeper understanding of the change, it can be viewed as a necessity to be able capture the whole picture of the actual change within the patient. The qualitative aspects are not a part of the statistically possible outcomes due to the lack of ability of adding the value of words in context of meaningful change. What this kind of model potentially can do is adding this aspect to create a more correct assessment of, for example, a treatment. An envisioned utilization of the method so far, is that it potentially can contribute with an approach to study meaningful change. For example, in different contexts and identifying which aspects that are likely to be seen as meaningful or not. It could be seen as a

method to understand meaningful change better, and it could be done by asking patients and clinicians to rate this and compare what they think is important.

Future potential of utilization could, for example, be the way it can be seen as a complement to other change measures, as implicated being needed by Blampied (2022). It could also potentially help psychologists to easier follow up their patients after a treatment. Suppose a psychologist maintains a website where patients have individual accounts to track their personal progress. Following the completion of treatment, patients are requested to provide monthly written descriptions of their mental health over a period of six months. By analysing the selected words, the meaningful change model can provide an estimate of the meaningful change. For instance, the psychologist may receive a notification from the model indicating a "relapse into depression" or an "indication of sustained improvement in mental state post-treatment." These notifications serve to expedite intervention in cases of deterioration or act as an efficient method of monitoring when patients are not deteriorating and may even be progressing towards a better mental state than before treatment. The psychologist could also potentially use the model as a complementary tool for confirming its own assessment of the patient, seeing if they correlate. It is important to highlight that AI is not intending to replace the human contact between patient and psychologist.

Limitations and future research

While this study has provided indications of validation and reliability, there are still certain limitations that need to be acknowledged to ensure higher reliability. In this study, both meaningful change and valence in the change have been assessed and rated. However, due to the need for simplicity and time constraints, the model has been constructed solely based on the ratings of meaningful change assessments. This can be seen as a limitation in this research but presents an opportunity for future development of the model to explore whether incorporating valence can enhance its predictive capabilities. It is important to note that the different assessment scales do not contribute equally to the model, as the primary focus is on predicting meaningful change. Valence, on the other hand, can be viewed as a potential supplementary factor that captures positive and negative changes. In contrast, meaningful change solely assesses the presence and meaningfulness of the change, and since change can be in a positive or negative direction, it can potentially be valuable to incorporate valence in the model.

Another limitation worth noting is the significant moderate correlation observed in the predictions. To establish the feasibility of assessing meaningful change through human-rated assessment of natural language responses, further investigation into validity aspects and external validity is necessary. It is important to emphasize that our current testing has focused on a single type of reliability and validity. Therefore, additional investigations are required to comprehensively evaluate the reliability and validity of this assessment approach for clinical use. It is also important to consider the ethical and social implications associated with large language models. These models raise concerns related to bias, fairness, privacy, and the potential for malicious use. Understanding and addressing these implications is crucial for the responsible deployment and usage of large language models (Cheng et al., 2023; Kjell et al., 2023b).

Building upon the potential highlighted by Kjell et al. (2019; 2022; 2023b) in exploring language-based assessments, the findings of this study contribute a step towards demonstrating the feasibility of utilizing such assessments for measuring meaningful change. What is left is for further research to continue exploring this and hopefully reach the vision of being able to use an improved model based on this method in clinical settings. To facilitate this progress, it could potentially preferable be involving clinical experts in conducting human-rated assessments and comparing the results with various additional rating scales. Additionally, this study contributes to the proof of concept, as its successful implementation in our dataset suggests the possibility of training new or additional models in different expert groups or domains.

Conclusion

This study has examined to what extent language-based assessments can be used to improve the way meaningful change is assessed. The research findings provided evidence supporting the reliability of human-rated assessments in meaningful change, as indicated by the inter-rater reliability. The human-rated assessments compared to the PHQ-9 difference scores, also indicates evidence of validity in the specific human-rated assessments used in this study. The language-based assessment indicated to be more suitable for assessing meaningful change than the PHQ-9 difference scores. It also showed that the trained model would preferably be built on human-rated assessments compared to traditional rating scales (i.e., the PHQ-9) based on the result from the ridge regression.

These findings suggest that language-based assessments can be further developed to assess meaningful change, and preferably by including human-rated assessment.

References

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences, 120*(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

Blampied, N. (2022). Reliable change and the reliable change index: Still useful after all these years? *The Cognitive Behaviour Therapist, 15*, E50. https://doi.org/10.1017/S1754470X22000484

Bost, R. H., Wen, F. K., Basso, M. R., & Cates, G. R. (2008). Online tools for evaluating patient change: Statistical foundations, clinical applications, research relevance. *Rehabilitation Psychology, 53*(3), 313–320. https://doi.org/10.1037/a0012977

Byrom, B., Breedon, P., Tulkki-Wilke, R., & Platko, J. V. (2020). Meaningful change: Defining the interpretability of changes in endpoints derived from interactive and mHealth technologies in healthcare and clinical research. *Journal of Rehabilitation and Assistive Technologies Engineering, 7*, 1–8. https://doi.org/10.1177/2055668319892778

Cheng, S. W., Chang, C. W., Chang, W. J., Wang, H. W., Liang, C. S., Kishimoto, T., Chang, J. P., Kuo, J. S., Su, K. P. (2023). The now and future of ChatGPT and GPT in psychiatry. *Psychiatry and Clinical Neurosciences, 77*(11), 592–596. https://doi.org/10.1111/pcn.13588

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Devlin, J., Chang, M-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics, 1*, 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

Eisen, S.V., Ranganathan, G., Seal, P., & Spiro, A. (2007). Measuring Clinically Meaningful Change Following Mental Health Treatment. *The Journal of Behavioral Health Services & Research, 34*(3), 272–289. https://doi.org/10.1007/s11414-007-9066-2

Gruijters, S. L. K., & Peters, G.-J. Y. (2022). Meaningful change definitions: Sample size planning for experimental intervention research. *Psychology & Health, 37*(1), 1–16. https://doi.org/10.1080/08870446.2020.1841762

Guyatt, G., Walter, S., & Norman, G. (1987). Measuring change over time: assessing the usefulness of evaluative instruments. *Journal of Chronic Disease, 40*(2), 171–178. https://doi.org/10.1016/0021-9681(87)90069-5

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15*(4), 336–352. https://doi.org/10.1016/S0005-7894(84)80002-7

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*(1), 12–19. http://dx.doi.org/10.1037/0022-006X.59.1.12

Kjell, O., Giorgi, S., & Schwartz, H. A. (2023a). The text-package: An R-package for analyzing and visualizing human language using natural language processing and transformers. *Psychological Methods*. https://doi.org/10.1037/met0000542

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

Kjell, O. N. E., Kjell, K., & Schwartz, H. A. (2023b). Beyond Rating Scales: With Targeted Evaluation, Language Models are Poised for Psychological Assessment. *Psychiatry Research*. https://doi.org/10.1016/j.psychres.2023.115667

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,* 3918. https://doi.org/10.1038/s41598-022-07520-w

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

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. ArXiv, abs/1907.11692.

Matero, M., Hung, A., & Schwartz, H. A. (2021). Understanding RoBERTa's Mood: The Role of Contextual-Embeddings as User-Representations for Depression Prediction.

Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, 89 – 94. Preprint. https://doi.org/10.48550/arXiv.2112.13795

McAleavey, A. A. (2021). When (Not) to Rely on the Reliable Change Index. https://doi.org/10.31219/osf.io/3kthg

McHorney, C. A., & Tarlov, A. R. (1995). Individual-patient monitoring in clinical practice: are available health status surveys adequate?. *Quality of Life Research, 4*(4), 293–307. https://doi.org/10.1007/BF01593882

Otsuka, N., Kawanishi, Y., Doi, F., Takeda, T., Okumura, K., Yamauchi, T., Yada, S., Wakamiya, S., Aramaki, E., & Makinodan, M. (2023). Diagnosing psychiatric disorders from history of present illness using a large-scale linguistic model. *Psychiatry and Clinical Neurosciences, 77*(11), 597–604. https://doi.org/10.1111/pcn.13580

R Core Team. (2023). *_R: A Language and Environment for Statistical Computing_*. R Foundation for Statistical Computing. https://www.R-project.org/

Schmidt, S., & Power, M. (2005). Clinical Psychology. *Encyclopedia of Social Measurement, 1*, 309–315. https://doi.org/10.1016/B0-12-369398-5/00513-2

Sikström, S., Pålsson, H. A., & Kjell, O. (2023). Precise language responses versus easy rating scales–Comparing respondents' views with clinicians' belief of the respondent's views. *PLOS ONE, 18*(2), e0267995. https://doi.org/10.1371/journal.pone.0267995

Uher, J. (2023). What's wrong with rating scales? Psychology's replication and confidence crisis cannot be solved without transparency in data generation. *Social and Personality Psychology Compass, 17*(5), e12740. https://doi.org/10.1111/spc3.12740

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1706.03762

Wolters Kluwer Health. (2000). Glossary: Health Outcomes Methodology. *Medical Care, 38*(9), II7–II13. http://www.jstor.org/stable/3768059