Natural Language Response Formats for Assessing Depression and Worry with Large Language Models: A Sequential Evaluation with Model Pre-registration

Zhuojun Gu[1], Katarina Kjell[1], H. Andrew Schwartz[2], and Oscar Kjell[1,2]

[1] Department of Psychology, Lund University
[2] Department of Computer Science, Stony Brook University

Large Language Models can transform individuals' mental health descriptions into scores that correlate with rating scales approaching theoretical upper limits. However, such analyses have combined word- and text-responses with little known about their differences. We develop response formats ranging from closed-ended to open-ended: 1) select words from lists, write 2) descriptive words, 3) phrases, or 4) texts. Participants answered questions about their depression/worry using the response formats and related rating scales. Language responses were transformed into word embeddings and trained to rating scales. We compare the validity (concurrent, incremental, face, discriminant, and external validity) and reliability (prospective sample and test-retest reliability) of the response formats. Using the *Sequential Evaluation with Model Pre-registration* design, machine-learning models were trained on a development dataset (*N*=963), and then *pre-registered* before tested on a prospective sample (*N*=145). The pre-registered models demonstrate strong validity and reliability, yielding high accuracy in the prospective sample (*r=.60–.79*). Additionally, the models demonstrated external validity to self-reported sick-leave/healthcare visits, where the text-format yielded the strongest correlations (being higher/equal to rating scales for 9 of 12 cases). The overall high validity and reliability across formats suggest the possibility of choosing formats according to clinical needs.

Keywords. Artificial intelligence, large language models, natural language, natural language processing, psychological assessment, depression, anxiety.

Acknowledgement. We thank Veerle Eijbroek for her thoughtful feedback on earlier drafts. Zhuojun Gu was funded by a grant from eSSENCE@LU (7:5) and the Department of Psychology at Lund University, and Oscar Kjell was funded by Marianne och Marcus Wallenbergs stiftelse (MMW 2021.0058), Katarina Kjell was funded by FORTE (2022-01022). 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. H. Andrew Schwartz was supported in part by a grant from the US CDC/NIOSH (U01 OH012476) and a grant from the NIH-NIAAA (R01 AA028032).

Conflict of Interest. O.N.E. Kjell and K. Kjell co-founded a start-up that uses computational language assessments to diagnose mental health problems. Zhuojun Gu and H. Andrew Schwartz report having no conflicts of interest with respect to the contents, authorship, or publication of this article.

Ethical approval. The Swedish National Ethics Review Board has deemed this research study (Ethics Application 2020-00730) exempt from requiring ethical approval according to Swedish

Law (see §§ 3-4 of the Act [2003:460] on ethical review of research involving humans in Sweden).

Recent advances in Artificial Intelligence, including Natural Language Processing and Deep Learning, have substantially improved the analysis of textual data (Bommasani et al., 2021; Vaswani et al., 2017) and the ability to assess psychological aspects of language (e.g., Boyd & Schwartz, 2021; Demszky et al., 2023; O.N.E Kjell, K. Kjell & Schwartz, 2024). AI-based language analyses have successfully been used in research for psychological assessments of, for example, personality (converging with rating scales; Cutler, & Condon, 2023; Kwantes et al., 2016; Schwartz et al., 2013), depression (from medical records, Eichstaedt et al., 2018, and Vaci et al., 2020; assessed with rating scales in Matero et al., 2021), and anxiety (assessed with rating scales, Rutowski et al., 2020). The types of language that have been used for computational language assessments have ranged from social media texts (e.g., from Facebook [Marengo et al., 2021; Park et al., 2015] and Twitter [Coppersmith et al., 2014; Eichstaedt et al., 2015]), audio recordings of everyday conversations (Gumus et al., 2023; Sun et al., 2020), and open-ended responses to prompted questions (O.N.E. Kjell et al., 2019). The latter approach recently yielded the highest accuracy in converging with rating scales, where respondents described their state of mind with descriptive words and text responses (O.N.E Kjell et al., 2022). Combining responses of different formats (words and text responses) across different construct questions (responses about harmony in life and satisfaction with life) was used to assess well-being rating scales with an accuracy (r = .85, N = 608) approaching the rating scales' own reliability (r = .71–.84, N = 608), which is considered a theoretical upper limit of assessment accuracy (O.N.E. Kjell et al., 2022).

The overall aim of this study is to develop and evaluate open-ended response formats for assessing mental states, with a focus on enhancing clinical utility. Mental health assessments are critical for effective diagnosis, treatment planning, and monitoring progress (Hunsley & Mash, 2007; Shen et al., 2018; Stahnke, 2021). Current methods, such as rating scales, are limited in their ability to comprehensively capture the complexity of patients' unique subjective experiences (Fried, 2017; Kjell et al., 2024), which can lead to missed nuances and suboptimal treatment outcomes. To address these challenges, this study develops and evaluates different response formats ranging from closed to more open, including 1) selecting descriptive words from pre-defined word lists (the select words format) and *writing* their own, 2) descriptive words (the write words format), 3) descriptive phrases (the write phrases format), or 4) descriptive texts (the write texts format).

The focus is on the two most common mental health conditions: depression and anxiety/worry, where self-reported subjective and behavioral experiences (e.g., low mood, worry, sleep disturbances) are key symptoms. Our selection of response formats offer varying levels of openness to describe relevant experiences, enabling a tailored approach to assessment that can potentially balance the need for brevity with the richness of open-ended responses. By exploring these formats, this study aims to provide clinicians and researchers with tools that can achieve high accuracy and reliability as well as offer more comprehensive insights into patients' mental health experiences.

Current Clinical Assessment Practices and Limitations

Current clinical practices primarily rely on a combination of closed-ended rating scales and unstructured clinical interviews to assess mental health problems. In an international survey of clinicians from the US (N = 148), the UK (N = 162), the Netherlands (N = 105), and Sweden (N = 127), revealed that clinicians working with mental health assessments primarily use unstructured clinical interviews (83%) and rating scales (83%), followed by semi-structured

interviews (65%) and structured interviews (40%; Navandi et al., in preparation). The survey also highlighted that mental health assessments require a significant investment of clinicians' time, with a median reported assessment time of 60 minutes, underscoring the time-intensive nature of current practices.

Unstructured interviews allow clinicians to explore patients' mental health in a more open-ended and conversational manner. However, these interviews are often time-intensive, require significant clinical expertise, and their reliability can vary widely (Regier, et al., 2013). Semi-structured and structured interviews offer greater standardisation and reliability, but they remain time-consuming and may still rely on the clinician's interpretation for nuanced responses (Hoffman et al., 2024; Lundqvist et al., 2022). The validity of clinical judgment is also compromised by cognitive biases and factors that increase susceptibility to errors, such as clinician fatigue (Bowes et al., 2020; Saposnik et al., 2016).

The quantification of psychological assessments has typically been achieved using closed-ended rating scales—based on questions or statements coupled with a limited set of pre-defined, categorical, or rating scale-based response formats (e.g., Likert, 1932). The closed-ended format provides consistency through predefined questions and response options, which are easy to administer. However, the closed-ended format of rating scales inherently limits the comprehensiveness and nuance of the information they can capture, often reducing patients' complex mental health experiences into closed-ended, forced-choice responses (Kjell et al., 2024). This closed-ended format restricts respondents from fully expressing their unique symptoms and experiences (Kjell et al., 2019; Pennebaker & Beall, 1986). Moreover, the closed-ended format comes with the risk that the scale developers potentially miss important aspects to capture a diagnosis or psychological construct (Fried et al., 2017; Kjell, 2011).

These limitations emphasise the need for more open-ended assessment methods that are also reliable in capturing the subjective and multifaceted nature of mental health problems. Open-ended language-based assessments offer a promising alternative, leveraging natural language responses to enable individuals to describe their unique experiences while producing reliable measurements with high accuracy.

Natural Language as the Foundation of Psychological Assessments

Language is at the core of the communication between patients and clinicians in clinical settings, playing a key role in assessment, treatment as well as evaluation of progress (Cruz & Pincus, 2002; Greenberg & Pascual‐Leone, 2006; Levinson & Holler, 2014). An open-ended approach using natural language as a foundation for quantitatively assessing mental states has many potential benefits. First, natural language is the natural way of communicating complex psychological states of mind (e.g., Tausczik & Pennebaker, 2010). Second, natural language comprises great measurement characteristics, including high range, resolution, dimensionality, and openness (O.N.E. Kjell et al., 2024). The range in language facilitates the description of extremes (e.g., *despondent* and *euphoric*). The resolution in language enables descriptions of detailed nuances (e.g., *despondent*, *sad*, *happy*, *euphoric*). The multi-dimensionality of language facilitates capturing complex states that vary in more than one-dimension (e.g., both *valence* and *arousal*). The openness of language affords the option to create personal responses, which is very difficult to capture exhaustively with predefined alternatives.

Third, the openness of language can enhance the content validity of mental health assessments by capturing a broader and more comprehensive representation of clinical constructs (e.g., depressive disorders). In contrast, the omission of important response alternatives in closed-ended methods can undermine construct validity by failing to account for the full breadth and complexity of the targeted disorder. For example, *mental pain* is commonly reported as a significant aspect of depression. To identify symptoms that matter most to patients, Chevance et al. (2020) asked participants to describe "the most difficult aspect of depression to live with or endure." They found that "mental pain" was the most frequently mentioned symptom. Yet, it is notably absent from the DSM-5 criteria and the closed-ended rating scales commonly used to assess depression, including those in this study. In previous studies using language-based assessments, descriptive words such as "painful" emerged as significantly related to high depression (Kjell et al., 2021). This highlights the limitations of predefined response categories and the potential of open-ended responses in capturing the comprehensive spectrum of patients' experiences.

Fourth, language-based responses also carry more information than closed-ended responses. Based on the information theory (Shannon, 1948), research has shown that natural language-based responses carry 4.8 times more information than closed-ended rating scales (O.N.E. Kjell et al., 2024), offering richer data. These favourable measurement characteristics indicate that using language-based responses has the potential to improve mental health assessments. This study focuses on comparing the characteristics of different response formats.

Different response formats may differ in multiple relevant dimensions for assessment, such as the type and amount of information they capture, the assessment accuracy they produce, the way responses can be visualised, and the time taken for patients to complete them. Additionally, asking respondents to answer the same question using different response formats may encourage more thorough answers, as each format may elicit different perspectives and capture unique information. Such an approach aligns with the cognitive interview theory (Willis, 2004), which emphasises how different questions can activate diverse cognitive processes, leading to the retrieval and articulation of different facets of an individual's experience, resulting in a more comprehensive answer. Hence, the response formats might complement each other, potentially enhancing the overall accuracy of a mental health assessment when combined (i.e., incremental validity).

Computational Language Assessments

Computational language assessments based on individuals' naturally occurring language, such as their Facebook statuses and tweets, have been analyzed with AI to assess psychological constructs. For example, Facebook statuses have been analyzed to assess individuals' Big Five personality factors (r = .54 - .63, N = 1943 in Lynn et al., 2020; r = .30 - .46, N = 4824 in Park et al., 2015; r = .28 - .42, N = 71,000 - 75,000 in Schwartz et al., 2013), depression levels (r = .39, N = 1000; Schwartz et al., 2014), and well-being (r = .20 - .30, N = 2198; Schwartz et al., 2016; see also Kjell et al., 2023). Facebook statuses have also been used to predict depression diagnosis from medical records with an accuracy of AUC = .72 (N = 683; Eichstaedt et al., 2018), where an AUC of .70 is considered the threshold for "sufficient" discrimination (Mokkink et al., 2018). However, these social media analyses rely on language that has not been directly prompted by specific questions, which is a key difference from this study.

Probed-based computational language assessments, where respondents describe their answers using natural language that is analyzed with AI, have shown promise in assessing psychological constructs with high validity. Computational language assessments where respondents describe their depression, worry, harmony in life, and satisfaction with life using descriptive words have been demonstrated to *measure* the degree of the constructs (i.e., concurrent validity; O.N.E. Kjell et al., 2019). Computational language assessments have further been shown to *describe* and *differentiate* well between similar constructs; for example, statistically significant words describe and differentiate well between depression versus worry, where, when compared with each other, depression is related with theoretically relevant words such as *sad*, *worthless*, and *lonely*, and worry with words such as *anxious*, *nervous*, and *scared*. Hence, these analyses have shown that language-based assessments have high face validity; here, we will examine how well the different response formats describe the different constructs and their face validity.

The Response Format

To date, respondents have been asked to answer the questions using descriptive words or descriptive text (O.N.E Kjell et al., 2019). Initially, the answers were analyzed using a bag-of-words approach called Latent Semantic Analyses (Landauer & Dumais, 1997) – where words are represented with a numeric representation that does not take the order and context of a word into account. This bag-of-words technique produced considerably more accurate results with the descriptive words (r = .34-.72, N = 91-852) than with the descriptive text format (r = .19-.49, N = 92-689; O.N.E. Kjell et al., 2019). A re-analysis of the data, with a *large language model* (based on a technique called *transformers*; Vaswani et al., 2017) – that takes words' order and context into account – has produced high accuracy for both descriptive words (r = .66 - .79, N = 608; O.N.E. Kjell et al., 2022) and descriptive texts (r = .61 - .74, N = 608; O.N.E. Kjell et al., 2022). The research has also demonstrated that combining the two response formats for two different constructs produces even higher accuracies (incremental validity; r = .85, p < .001, N = 608) that rival the rating scales' own reliability as measured with test-retest correlation, corrected item-total correlation average, and inter-item correlation average (r = .71 -.84, N = 608). This accuracy is an essential achievement because the scale's own reliability can be seen as a theoretical upper limit of predicting it. Hence, the large language models have opened up the possibility of using more complex response formats while attaining high measurement accuracy.

Large Language Models in Natural Language Processing (NLP)

Large language models are now the foundation for most state-of-the-art AI-language systems and have been described as a "paradigm shift" (Bommasani et al., 2021, *p*. 1). Large language models can efficiently represent a word's meaning based on the context in which they were used. The context a word is in *plays* a vital role in determining its meaning. For example, consider how the context changes the meaning of the word *play* in the previous sentence, as compared with "I *play* my guitar" versus "I watched a *play*". This technique has led to increased accuracy in standard Natural Language Processing tasks such as grammatical acceptability (Warstadt et al., 2019) and question answering (Rajpurkar et al., 2018). The ability of large language models to more accurately represent language may provide more options for using different response formats. Whereas previous computational language assessments have found the highest assessment accuracy of the write words response format, the advent of large language models that more accurately account for grammar and syntax, including word order, may provide even higher accuracy than those previously achieved.

Depression and Anxiety

Here, we focus on assessing depression and anxiety for three primary reasons. First, depression and anxiety are highly prevalent in the general population, representing two of the most common mental health conditions globally (Baxter et al., 2013; Lépine & Briley, 2011). Given their high prevalence, improving assessment methods for these conditions has significant utility. Second, the assessment of depression and anxiety heavily relies on subjective experiences, as these disorders are characterised by internal states such as low mood and worry (e.g., see the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, DSM-5; American Psychiatric Association, 2013). Open-ended language-based assessments provide an opportunity to more comprehensively quantify these subjective experiences. Last, previous research using language-based assessments shows that writing words to describe one's level of depression and anxiety significantly converges with the individual item scores and total scores of corresponding rating scales (K., Kjell et al, 2021). In this study, we build on these insights to examine whether different response formats can further improve the assessment validity and reliability of depression and anxiety measures.

Research Questions and Hypotheses

We evaluate the four response formats: 1) the select words, 2) the write words, 3) the write phrases, and 4) the write texts response format according to their:

Validity of models during development

i) **Concurrent (criterion) validity.** How well do they converge with rating scale scores?
ii) **Concurrent (criterion) validity across sample sizes.** To what extent does the model accuracy increase with larger training sample sizes?
iii) **Incremental validity.** How well do combinations of response formats converge incrementally with rating scales?
iv) **Face validity.** Is the language that is predictive of depression versus anxiety face valid (when depicted in word vizualisations)?
v) **Discriminant validity.** How well do they differentiate between depression and anxiety?

Validity and reliability of pre-registered models

vi) **Prospective sample reliability.** How accurate (convergence with rating scales) are pre-registered models on data from a new set of participants?

This part involved two pre-registered hypotheses stating that the pre-registered models of depression and anxiety/worry yield scores that:

a) are positively associated with corresponding rating scale scores, and
b) achieve at least a moderate correlation $r > .50$ for total scores where depression language predicts depression rating scales and worry language predicts worry/anxiety rating scale scores (note that the 99% confidence interval for the lowest correlation in the training set [$r = .59$] is above .50 [.53 – .64 $N = 963$]).

vii) **Test re-test reliability in a prospective sample.** How high is the two-week test-retest correlation for the pre-registered models?

viii) **External validity in a prospective sample.** How well do the pre-registered models' predictions correlate with a few self-reported behaviors associated with depression and anxiety (e.g., sick leave due to mental health issues)?

Descriptive characteristics of the response formats

ix) **Information content.** How much information content is captured by each response format?

x) **Time burden.** What is the response time for each format?

Methodological Disclosure and Data Accessibility Statement

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. We make all the research materials, and the anonymised prospective data, publicly available anonymously at the OSF (https://osf.io/syren/?view_only=9eebf3e448b447ae89dd288d3107cb03).

Methods

We use the *Sequential Evaluation with Model Pre-registration* (SEMP; O.N.E. Kjell et al., 2024) study design, which aspires to adhere to good scientific practices (i.e., pre-registration) combined with development practices for achieving robust AI model evaluation (e.g., out-of-sample testing). Here, SEMP involves registering models developed in a *development set* ($N = 963$) and pre-registering hypotheses before testing them on a new *prospective participant set* ($N = 150$).

Participants

The development data: Participants were recruited online using Prolific (Palan & Schitter, 2018) in August 2020. A mixture of unscreened and screened respondents were recruited. Screened participants reported being diagnosed with an ongoing mental health diagnosis of Major Depressive Disorder (MDD) or Generalized Anxiety Disorder (GAD). Overall, 258 participants reported having been diagnosed with MDD, 259 with GAD, and 491 were recruited from an unscreened population. Forty-five participants were excluded for not answering the attention check items correctly ($N = 32$) or not having any data saved ($N = 13$). Out of the remaining 963 participants, 571 identified as female, 388 as male, two preferred not to say, and two did not respond. The average age was 33.3 ($SD = 11.2$; *range* = 18 – 77) years. Participants reported being from the U.S. (n = 284) and the U.K. (n = 679), and 903 reported English, 16 reported European languages, 16 reported Asian languages, 1 reported Russian, 3 reported African languages, and 1 reported Esperanto as their first languages, with 23 participants not responding[1]. The training dataset sample size was determined based on previous studies (O.N.E. Kjell et al., 2019; 2016). In the training set, 7 participants had missing data in the write text response format questions (7 missing for the depression and 3 missing for the worry questions).

[1] One participant used words like "Yes, No, No, No, No" in "word" format to respond. Another participant did not respond to "word", "phrase" and "text". Both cases are included in the results given that the number is small relative to the sample size, and the unintended responses occurred only in restricted responses.

The prospective test data: Participants (N = 150) in the prospective test data were also collected from Prolific in January 2024 (i.e., collected 41 months after the development set). The size was based on finding reliable estimates of the correlations that we were interested in from analyses of the training dataset. Five participants were excluded from the analyses: three did not pass the attention check items, and two did not respond to the write text questions of depression and anxiety[2]. The rest of the participants (N = 145) were further invited to take part in a longitudinal study with two weeks between Time 1 and 2 (T1, T2). At T1, participants had a mean age of 41.6 (SD = 12.6, *range* = 19 – 81) years; 101 female, 43 male, and 1 other. Their reported nationality included 126 from the U.K., 5 from the U.S., and 14 from other countries. At T2, 122 (84.1%) participants returned to answer the same survey again; they have an average age of 42.3 (SD = 12.6; *range* = 19 – 81) years, and 87 identified as female, 34 as male, and 1 as other. All participants in the prospective test data have English as their first language.

Measures

Screening Question

The participants were asked to report whether they had any previous or ongoing diagnosis of MDD, GAD, or other mental health problems (see Supplement Appendix (SA) 1 for details).

The Response Format Questions

The open-ended response format for descriptive words (developed by O.N.E Kjell et al., 2019) was used and adapted for the different response formats. Previous research piloted different question prompts and instructions, finding that this format with a question prompt coupled with instructions to elaborate using open-ended responses works very well (Kjell et al., 2019). An example question of the write phrases and write text formats are exemplified below:

> **Over the last 2 weeks, have you been depressed or not?**
>
> Please answer the question by typing 5 descriptive words or phrases below that indicate whether you have been depressed or not. For example, if you have been depressed, then write more and stronger words or phrases describing this, and if you have not been depressed, then write more and stronger words or phrases describing that.
>
> Write descriptive words or phrases relating to those aspects that are most important and meaningful to you.
>
> Write <u>one to five</u> words in each box.

> **Over the last 2 weeks, have you been depressed or not?**
>
> Please answer the question by typing at least a paragraph below that indicates whether you have been depressed or not. Try to weigh the strength and the number of aspects that describe if you have been depressed or not so that they reflect your overall personal state of depression. For example, if you have been depressed, then write more about aspects describing this, and if you have not been depressed, then write more about aspects describing that.

[2] The pre-registration did not include these exclusion steps, they follow previous research; e.g., O.N.E Kjell et al., 2016; 2022.

Write about those aspects that are most important and meaningful to you.

Write at least one paragraph in the box.

The question was adapted for worry ("worried or not"), and the instructions were adapted for the select words ("answer the question by selecting 5 descriptive words"), and the write word formats ("answer the question by typing 5 descriptive words"; see SA2). The list of words in the select words format, comprised the 31 most frequent words participants used to describe their depression (e.g., *blue*, *lonely*, *tired, hopeful*, *happy*) or worry (e.g., *anxious*, *nervous*, *worried*, *calm*, *peaceful*) in previous research (O.N.E. Kjell et a., 2019; see SA3 for the lists).

Rating scales

Several commonly used clinical scales are used to measure depression and anxiety severity, with the aim for the scales to complement each other. The Patient Health Questionnaire-9 for depression (PHQ-9, Spitzer et al., 1999) and the Generalized Anxiety Disorder-7 for anxiety (GAD-7, Spitzer et al., 2006) are closely aligned with the symptoms defined in the DSM manual for Major Depressive Disorder and Generalized Anxiety Disorder, respectively. The Center for Epidemiologic Studies Depression Scale (CES-D, Radloff, 1977) complements the PHQ-9 by including some depression-related symptoms not included in the DSM, such as indirectly related interpersonal issues. The Penn State Worry Questionnaire (PSWQ, Meyer et al., 1990) complements the GAD-7 by focusing on experiences of worry.

The Patient Health Questionnaire-9 (PHQ-9). The PHQ-9 (Spitzer et al., 1999) measures major depressive disorder as defined in the DSM-IV. The rating scale comprises nine items addressing symptoms such as "Little interest or pleasure in doing things". The scale uses a four-point scale (0 = *Not at all* to 3 = *Nearly every day*). The items refer to experiences from the last two weeks. In the current study, Cronbach's alpha was .95, and MacDonald's hierarchical omega was .84.

The Center for Epidemiological Studies Depression Scale (CES-D). The CES-D measures depressive symptoms severity (Radloff et al., 1977). It includes 20 items referring to the last week, such as "I could not get 'going'", coupled with a four-point scale (0 = *rarely or none of the time [less than one day]* to 3 = *most or all of the time [10-14 days]*). Cronbach's alpha was .96, and MacDonald's hierarchical omega was .91 in this study.

The Generalized Anxiety Disorder-7 (GAD-7). The GAD-7 measures Generalized Anxiety Disorder as described in the DSM-IV (Spitzer et al., 2006). It comprises seven items, for example, "Feeling nervous, anxious or on edge", referring to the last two weeks, coupled with a four-point Likert scale (0 = *Not at all*, to 3 = *Nearly every day*). Cronbach's alpha was .93, and MacDonald's hierarchical omega was .87 in this study.

The Penn State Worry Questionnaire (PSWQ). The PSWQ was developed to assess worry/anxiety symptoms severity (Meyer et al., 1990). The rating scale comprises 16 items, such as "My worries overwhelm me". The scale uses a 5-point Likert scale (0 = *Not at all typical of me*, 4 = *Very typical of me*). In this study, Cronbach's alpha was .91, and MacDonald's hierarchical omega was .89.

Demographics and mental health-related sick leave and healthcare visits

The survey included asking participants to report their gender (including *Female, Male,* or *Other*), age, sick leaves (due to mental health issues) in the last three months and the last year, and healthcare visits (due to mental health issues) in the last year.

Attention check items

Attention check items instructing participants to select a specific rating scale response option (see SA4 for details) were included within the PHQ-9 and the GAD-7.

Procedure

All participants were informed about the nature of the study, that their participation was voluntary, that they could withdraw at any time without having to give a reason, and that their answers were anonymous. All participants provided their consent to take part and that their anonymized data could be shared openly. Then, participants are asked whether they previously had been diagnosed with MDD, GAD, or any other mental health problems. After that, the survey had five parts: First, participants answered the three open-ended response formats (i.e., write words, phrases, and texts), which were presented in random order across participants. Second, they answered the select words questions. Third, they filled out the rating scales: the PHQ-9, the CES-D, the GAD-7, and the PSWQ, which were randomized. Last, they reported the demographics and mental health-related sick-leave and healthcare visits. Participants completed the study in a median time of 20.1 (*mean* = 24.0, *SD* = 12.8) minutes.

Analytical methods

The analyses were conducted in R (R Core Team, 2018) and the text-package (version 1.2.1; O.N.E. Kjell et al., 2023). The text-package enables social scientists to access open large language models, and machine learning and natural language processing methods. We employed large language models to convert language responses into word embeddings (i.e., numerical representation). These embeddings were then used as predictors in training regression-based machine-learning techniques to predict scale scores.

Pre-trained word embeddings

We use a large language model called RoBERTa-large (Liu et al., 2019), which has been developed from text from Wikipedia and books. The RoBERTa-Large model has demonstrated strong alignment between language-based assessments and a range of psychological constructs, incorporating both self-reports and expert evaluations (Ganesan et al., 2021; Matero et al., 2024; Varadarajan et al., 2024). We did not pre-process the text data; however, RoBERTa-large automatically tokenises text using the Byte-Pair Encoding (Sennrich, 2015). RoBERTa-large represents each word (token) across 24 layers, each with 1024 dimensions. Only the second-to-last layer is used here, as empirical research shows that these work well for representing the meaning of words when modeling human-level language tasks such as assessing mental health issues (Ganesan et al., 2021). Word embeddings are aggregated to represent several words/text using the means (the default in the text-package).

Predictive modeling using word embeddings

The word embedding dimensions of the responses are used as predictors in ridge regression (Hoerl & Kennard, 1970) to predict the rating scale scores. Although there are many types of predictive models from statistical learning where embeddings can be applied, recent empirical studies have consistently shown that ridge regression achieves state-of-the-art results when using

contextual embeddings (Ganesan et al., 2021; Matero et al., 2021; O.N.E. Kjell et al., 2022;). Tenfold cross-validation was used for model training. The dataset was split into a training set (90%) and a test set (10%). Within the training set, an analysis set was used for developing models with various penalties, and an assessment set was utilized to evaluate these penalties and select the optimal model for the testing set (for details, see O.N.E. Kjell et al., 2023). The search grid in the ridge regression covered the range of 10^{-16}–10^{16} with increases of times 10. The training sets were stratified into groups based on the outcome (y) using 4 bins. Pearson correlation between the observed and language-assessed scores is used to assess the accuracy. Appendix SA5 lists additional R packages involved.

Significance testing the assessment accuracy of models

To compare the errors of two prediction models, we compute the error for each prediction (i.e., $y - \hat{y}$), and then use a paired sample t-test to examine whether the errors significantly differ.

Standardised difference scores

To assess a form of discriminant validity of the different response formats, we evaluate their ability to converge with difference scores between rating scales. This is done by training word embeddings to evaluate the difference scores. The difference scores are computed by subtracting one normalized score from another (where the normalisation includes subtracting the mean and dividing by the standard deviation). For example, a positive score from taking the normalized PHQ-9 score minus the normalized GAD-7 score, indicates greater depressive symptoms *relative to* anxiety symptoms. This method allows us to determine how well the embeddings capture the distinctions (the relative difference) between these related constructs.

Diversity index

To quantify the information in responses, we used Shannon Entropy (Shannon, 1948). This measure is important in machine learning, as it indicates how much information the algorithms have at their disposal to learn from. In this article, we use the Diversity index, which quantifies the entropy or diversity in a set of values or probabilistic events, which is here the possible responses on a rating scale or a word-based response). Mathematically, the Diversity Index is computed as $2^{entropy}$, where *entropy* is calculated as the Shannon entropy:

$H = -\sum x \quad p(x) \log(p(x))$, where p is its probability mass function, and x is a set of values or probabilistic events.

The magnitude of the correlations

All correlations were computed using Pearson's correlation. Correlations of .20 to .39 were interpreted as weak, .40 to .59 as moderate, .60 to .79 as strong, and above .80 as very strong (O.N.E. Kjell et al., 2022).

Results

Descriptives

Table 1 presents the correlations among rating scales and their mean, median, and standard deviation. The correlations between the rating scales range from r = .64 - .91. The strong correlations indicate the convergent validity among the scales measuring depression and anxiety; notable though, the GAD-7 correlates stronger with the depression scales (r = .82-.85) than to the PSWQ (r = .64-.71).

Table 1.
Pearson Correlations and Descriptives of the Rating Scales

Measure	1.	2.	3.	4	5.	6.	Mean	Median	SD
The development data (*N* = 963)									
1. PHQ-9	-	-	-	-	-	-	11.56	11	7.56
2. CES-D	.91	-	-	-	-	-	26.84	28	14.95
3. GAD-7	.82	.85	-	-	-	-	10.10	10	6.27
4. PSWQ	.64	.71	.76	-	-	-	42.17	45	15.53
5. Sick-leave due to mental health last 3 months	.23	.24	.24	.15	-	-	4.38	0	17.05
6. Sick-leave due to mental health last year	.23	.24	.23	.16	.90	-	16.47	0	65.05
7. Healthcare visits due to mental health last year	.25	.26	.21	.16	.25	.23	.96	0	2.64
The prospective data (*N* = 145)									
1. PHQ-9	-	-	-	-	-	-	8.25	7	7.09
2. CES-D	.92	-	-	-	-	-	20.16	18	14.63
3. GAD-7	.82	.86	-	-	-	-	7.84	7	6.43
4. PSWQ	.65	.70	.76	-	-	-	37.88	41	17.33
5. Sick-leave due to mental health last 3 months	.29	.30	.21	.15	-	-	2.49	0	11.97
6. Sick-leave due to mental health last year	.22	.23	.19	.16	.77	-	6.41	0	32.71
7. Healthcare visits due to mental health last year	.35	.31	.19	.11	.62	.26	.48	0	1.78

Notes. PHQ-9: Patient Health Questionnaire - 9; CES-D: The Center for Epidemiological Studies Depression Scale; GAD-7: Generalized Anxiety Disorder - 7; PSWQ: Penn State Worry Questionnaire; Supplementary Material Table S1 has more descriptive information for the prospective data).

Concurrent and incremental validity

Using all eight response formats, including all four response formats for both depression and worry, yields the highest concurrent accuracies to rating scales, which approach theoretically upper limits (Table 2). Rating scales' reliability can be seen as a theoretical upper limit (Muchinsky, 1996) for assessment accuracy, which is here measured as the mean of the item-total correlation and the test-retest reliability. The language-based assessments converge with the PHQ-9 with a correlation of .78, which is close to its reliability (r = .79); and converges with the CES-D with a correlation of .83, which is higher than the scales' reliability (r = .78). Further, the language-based assessments converge with the GAD-7 with a correlation of .77, which is .05 lower than its reliability (r = .82), and converge with the PSWQ with a correlation of .74, which is .06 lower than its reliability (r = .80).

Table 2

Concurrent Validity: Comparing 10-fold Cross-Validated Pearson Correlations Based on Combined Responses with the Rating Scales' Reliability.

	Depression		Worry	
	PHQ-9	**CESD**	**GAD-7**	**PSWQ**
All 8 response formats	.78 (.76 - 1.00)	.83 (.81 - 1.00)	.77 (.74 - 1.00)	.74 (.71 - 1.00)
Mean reliability[1]	.79 (.74 - .82)	.78 (.73 - .81)	.82 (.77 - .85)	.80 (.76 - .84)

Notes. N = 963. All correlations were significant at $p < .001$. PHQ-9: Patient Health Questionnaire - 9; CES-D: The Center for Epidemiological Studies Depression Scale; GAD-7: Generalized Anxiety Disorder - 7; PSWQ: Penn State Worry Questionnaire.
[1] We are taking the average of the item-total correlation and the two-week test-retest reliability of the scales (see Table S5 for individual reliability scores. Results trained to log-transformed rating scale scores are presented in Table S7.

Concurrent validity of individual and combined response formats

Out of the four response formats, the select words response format yields the strongest assessment accuracy for depression (r = .73 for PHQ-9), followed by the write phrases (r = .69), the write text (r = .69), and the write words formats (r = .68; Table 3). In fact, the select format yields significantly lower errors than the write words (t = -5.00, $p < .001$), phrases (t = -4.11, $p < .001$), and texts (t = -3.70, $p < .001$). For worry, select words (r = .67 for GAD-7) and the write words (r = .67) formats performed the most accurately, followed by the write phrases (r = .62) and texts (r = .59) formats. Also, the select words format yields significantly lower errors than the write phrases (t = -3.61, $p < .001$) and texts (t = -4.37, $p < .001$) formats, but not than the write words (t = -.57, p = .566) format.

Notably, the language responses tend to show a consistent specificity across all the response format predictions, so *depression language* tends to assess *depression rating scales* more accurately than the anxiety/worry scales, and *worry language* assess the *anxiety/worry rating scales* more accurately than the depression rating scales.

Table 3.
Concurrent Validity: The 10-fold Cross-Validated Pearson Correlations of Single Format Model Assessments and the Observed Rating Scales

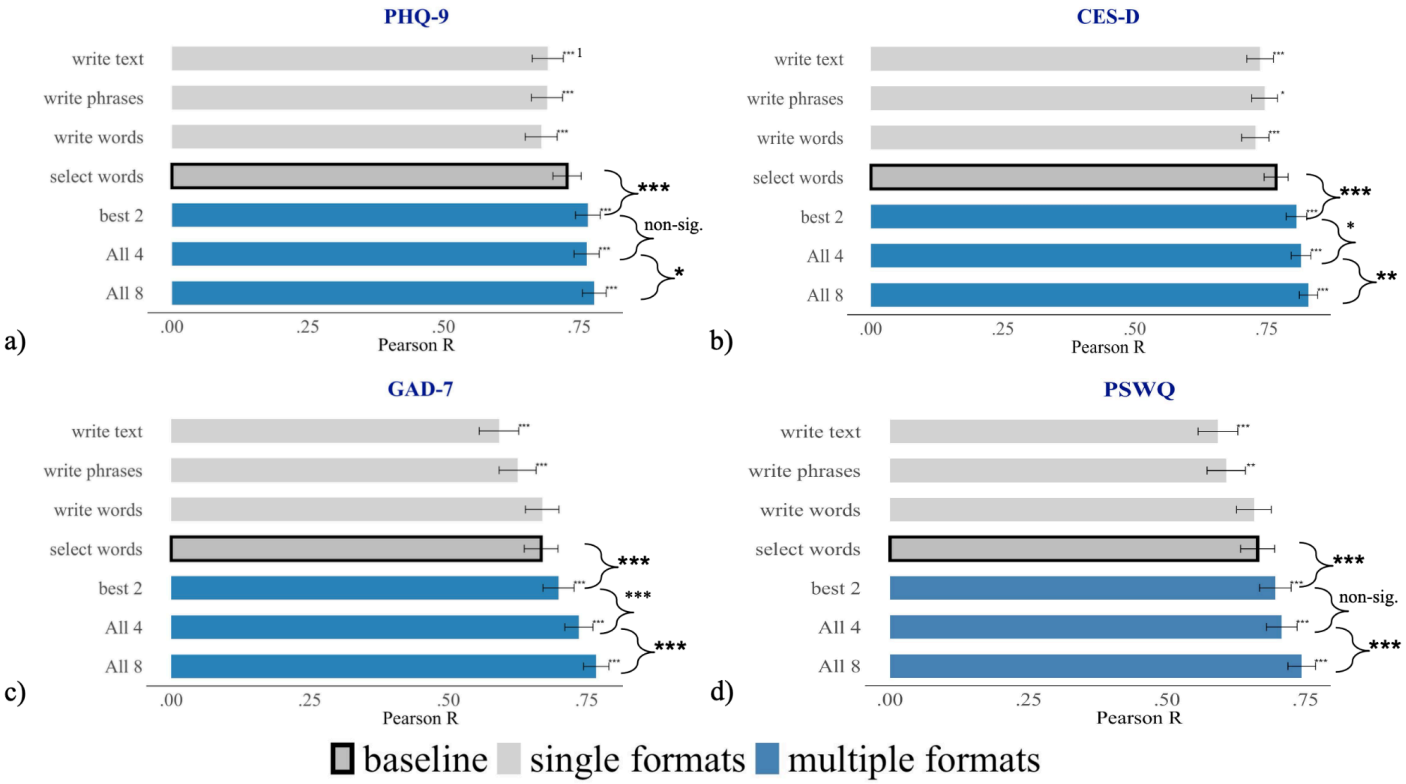
Response format	**Depression Prompt**		**Worry Prompt**	
	PHQ-9	**CES-D**	**GAD-7**	**PSWQ**
Select words	.73 (.70 - 1.00)	.77 (.74 - 1.00)	.67 (.64 - 1.00)	.66 (.63 - 1.00)
Write words	.68 (.65 - 1.00)	.73 (.70 - 1.00)	.67 (.64 - 1.00)	.66 (.63 - 1.00)
Write phrases	.69 (.66 - 1.00)	.75 (.72 - 1.00)	.62 (.59 - 1.00)	.61 (.57 - 1.00)
Write text	.69	.74	.59	.59

	(.66 - 1.00)	(.71 - 1.00)	(.56 - 1.00)	(.56 - 1.00)

Notes. *N* = *963.* All correlations were significant at $p < .001$.
PHQ-9 = Patient Health Questionnaire-9 assessing depression; CES-D = The Center for Epidemiological Studies Depression Scale (CES-D); GAD-7 = Generalized Anxiety Disorder - 7; PSWQ = Penn State Worry Questionnaire.
Depression prompt models for the PHQ-9 and the CES-D and worry prompt models for the GAD-7 and the PSWQ were pre-registered for the prospective data sample, which are presented in Tables 7, 8, and 9. For results where depression responses are trained to anxiety and worry scales and vice versa see Table S8.

Incremental validity: Combining response formats increases the assessment accuracy

Combining response formats tends to increase the assessment accuracy compared to only using one response format (Figure 1 and Table S11). For worry, combining all four formats yields the strongest correlation to the anxiety rating scales (r = .71-.74), whereas for depression, combining only the select words and the write text formats yields a stronger correlation to depression rating scales (r = .77-.81) than using all four formats (r = .76). Further, combining both response formats and construct questions improves the assessment accuracy further, as demonstrated when using all eight response formats, which produces the highest correlations (significance testing the errors for 4 [construct congruent models, where, e.g., depression language predict depression scales] versus 8 response format yields significant differences for the PHQ-9: $t = 2.26$, $p = .024$; the CES-D: $t = 2.73$, $p = .006$; the GAD-7: $t = 4.22$, $p < .001$; and the PSWQ: $t = 4.14$, $p < .001$. Significance testing the errors for 4 versus the 2 best formats yield significant differences for the CES-D ($t = 2.39$, $p = .017$), the GAD-7 ($t = 4.86$, $p < .001$), but not for the PHQ-9 ($t = -.98$, $p = .326$) and the PSWQ ($t = 1.24$, $p = .215$). Significance testing the errors for the 2 most accurate formats versus the select words format yields significant differences for the PHQ-9 ($t = 6.78$, $p < .001$), the CES-D ($t = 7.35$, $p < .001$), the GAD-7 ($t = 4.05$, $p < .001$), and the PSWQ ($t = 4.30$, $p < .001$).



[1] All the significant tests on the error bars are compared to select words.

Figure 1 | Incremental Validity: The Pearson Correlations of Single versus Combinations of Response Formats, a) the Patient Health Questionnaire (PHQ-9), b) The Center for Epidemiological Studies Depression Scale (CES-D), c) the Generalized Anxiety Disorder - 7 (GAD-7), d) the Penn State Worry Questionnaire (PSWQ).
All 8 = Eight response formats. All 4 = Four response formats with the same construct prompt. Best 2 comprise select words and write text responses from the same construct prompt, which all have the highest accuracy among the two response formats combinations. The error bars show the 95% of the confidence intervals of each correlation. The stars indicate which of the models that significantly differ in assessment error; *** = $p < .001$, ** = $p < .01$, * = $p < .05$. For more details see Table S11.

Concurrent validity across sample sizes

Assessing concurrent validity across sample sizes shows how much data is necessary to achieve reliable and accurate model performance in clinical assessments. Figure 2a-d shows the concurrent accuracy achieved by models trained on different sample sizes. Overall, the models based on the select words format, including the models based on 4 and all 8 response formats, reach a high convergent accuracy at around 100 participants (although with more participants, the accuracy increases and the error bars, of course, reduce). The writing-based formats require some more training examples than the select words format before reaching a more stable performance estimate; after 100 participants, all models are within ±.24 of the performance estimate based on all participants; after 300 participants, all models are within ±.09; after 500 participants, all models are within ±.05; and after 700 participants all models are within ±.03.

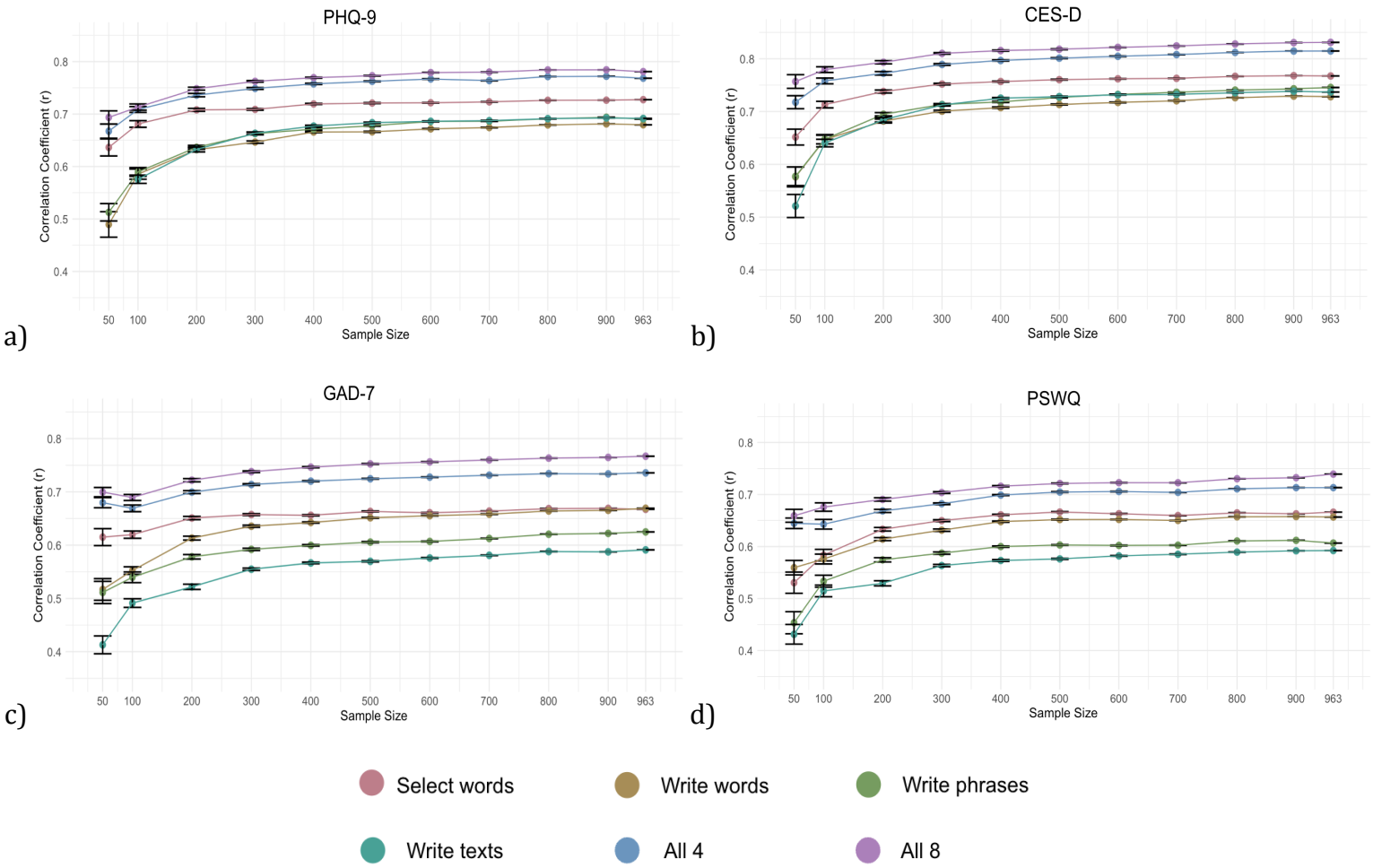


Figure 2 | Assessment accuracy (*Pearson r*) across the number of participants included in training for each response format and the combination of 4 and 8 formats assessing a) PHQ-9, b) CES-D, c) GAD-7, and d) PSWQ.

Face validity: Visualizations of the language predictive of depression and anxiety

Plotting statistically significant word responses according to rating scales shows words related to low versus high scores (Figure 3a-d). For all word plots, high levels of mental health are described with words such as *happy*, *glad*, and *blessed* – whereas high levels of depression tend to be described with words such as *sad* and *blue*, and worry with words such as *anxious*, *worried*, and *stress*. The select words format yields a more constrained descriptive representation (fewer statistically significant words) compared to the more open response formats. It is also noticeable that word plots based on text responses demonstrate more function words.

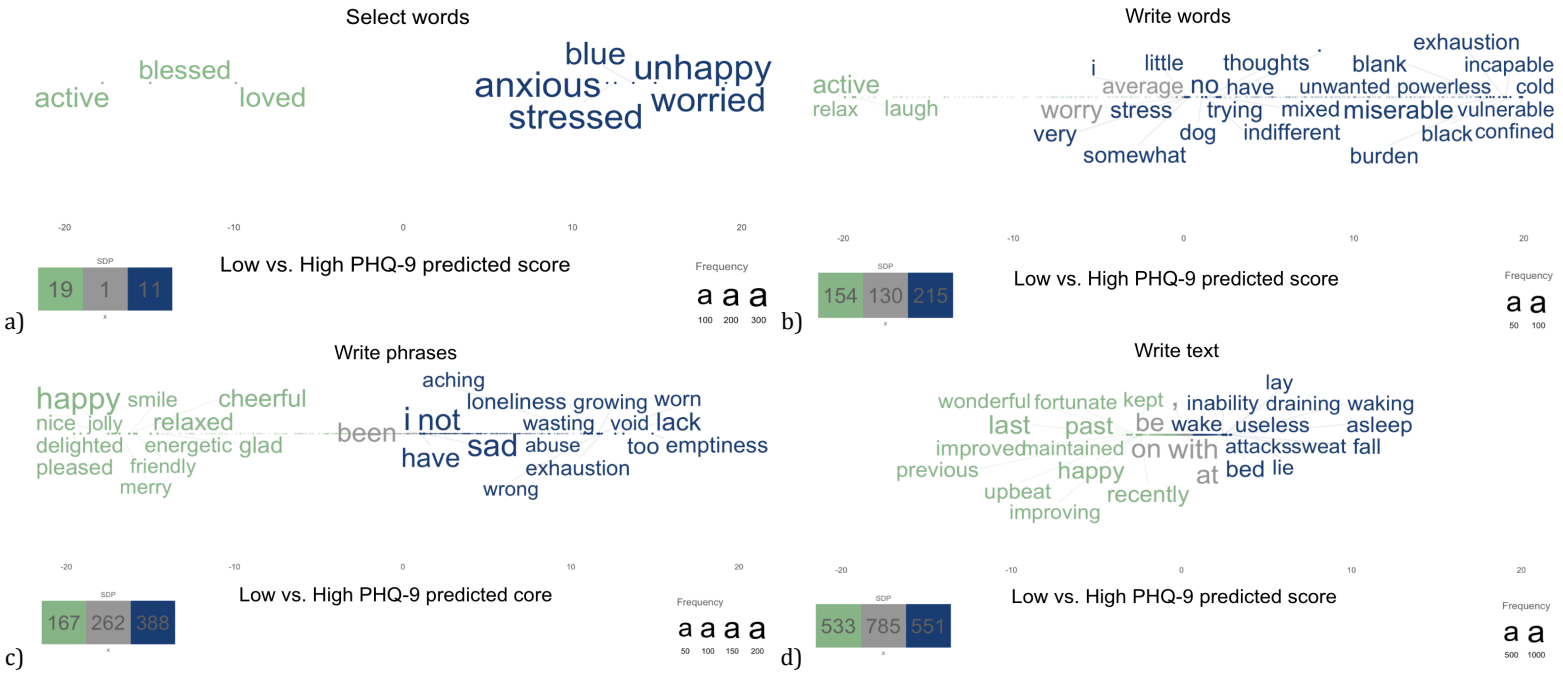


Figure 3 | Statistically significant words related to low (green) versus high (blue) predicted PHQ-9 scores across the response formats, including a) select words, b) write words, c) write phrases, and d) write text. The models were trained to the normalised rating scale scores. The more open the word responses are, the more statistically significant words. The open-ended response formats include more function words. Coloured words are statistically significant when correcting for multiple comparisons using “False Discovery Rate”**.**

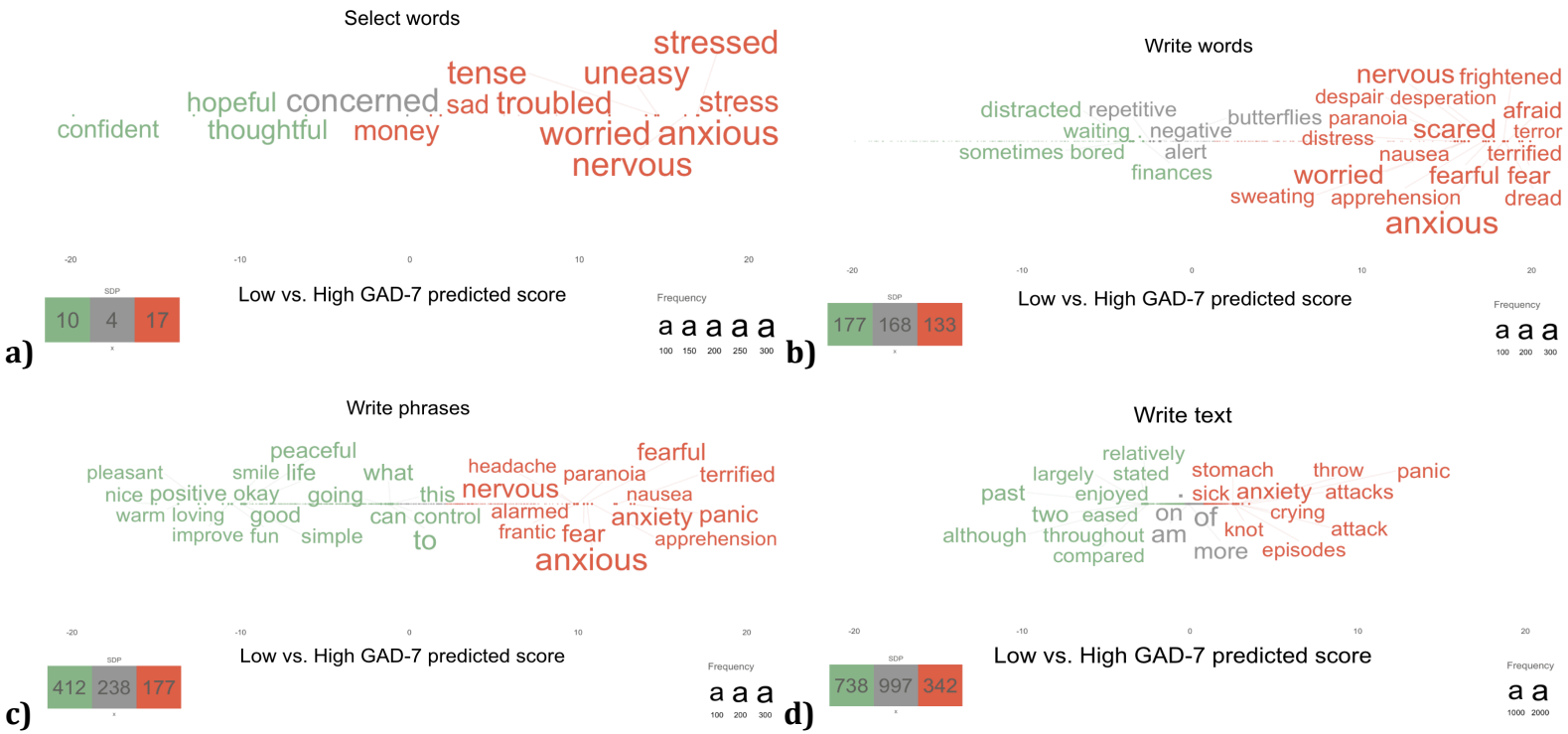


Figure 4 | Statistically significant words related to low (green) versus high (red) predicted GAD-7 scores across the response formats, including a) select words, b) write words, c) write phrases, and d) write text. The models were trained to the normalised rating scales scores. The more open the word responses are, the more statistically significant words there are. The open-ended response formats include more function words. Coloured words are statistically significant when correcting for multiple comparisons using “False Discovery Rate”.

Discriminant validity

Whereas models based on several response formats tend to yield the strongest correlations to a targeted rating scale (Tables 2, 3, S11), the model prediction scores of different rating scales yield extremely strong correlations (Table 4). For example, the correlations between the predictions of the depression rating scale the PHQ-9, and the predictions of the anxiety rating scale the GAD-7, based on all eight response formats correlate with $r = .93$, whereas when based on single response formats, the predictions only correlate with $r = .51 – .65$; this may indicate poor discriminant validity. Hence, using responses from more formats when training the model increases the assessment accuracy, which may be at the cost of discriminant validity.

Developing models assessing the difference score. Although using multiple response formats appears to yield predictions with lower discriminant validity (i.e., strong inter-correlations), it is possible to create models predicting the normalized difference score between rating scales more accurately with multiple, as compared with single response formats (Table 5; Table S3). With all eight formats, it is possible to assess the difference score of the PHQ-9 and the GAD-7 with a correlation of $r = .38$, whereas only using a single response format for one construct yields accuracies ranging from $r = .07$ to .20 (single responses in Table S3). Hence, it is possible to adapt the models according to different needs.

Visualizing differences between constructs. It is also possible to differentiate the words that statistically differentiate between closely related psychological constructs (Figure 5a-d). Statistically significant words that differentiate between depression and worry responses tend to reflect the core symptoms/criteria from the DSM-V (American Psychiatric Association, 2013). Depression responses are represented by words such as *blue*, *depressed,* and *emptiness* (c.f., DSM-V criteria depressed mood), and *meaningless* and *worthless* (c.f., DSM-V criteria diminished interest or pleasure). Worry responses are represented by words such as *worried, anxious, and nervous* (c.f., DSM-V criteria excessive anxiety and worry), *panic* (c.f., DSM-V criteria difficult to control the worry).

Table. 4

Discriminant Validity: Correlations between Language-Based Assessments Across Models based on Different Response Formats

	Language response format	Rating scale	1.	2.	3.
1.	**all 8**	**PHQ-9**	-	-	-
2.		**CES-D**	*.97*	-	-
3.		**GAD-7**	*.93*	*.95*	-
4.		**PSWQ**	*.85*	*.87*	*.93*
			5.	**6.**	**7.**
5.	**all 4 depression**	**PHQ-9**	-	-	-
6.		**CES-D**	*.98*	-	-
7.	**all 4 worry**	**GAD-7**	*.68*	*.68*	-
8.		**PSWQ**	*.64*	*.65*	*.93*
			9.	**10.**	**11.**
9.	**depression select words**	**PHQ-9**	-	-	-
10.		**CES-D**	*.99*	-	-
11.	**worry select words**	**GAD-7**	*.64*	*.65*	-
12.		**PSWQ**	*.62*	*.63*	*.97*
			13.	**14.**	**15.**
13.	**depression write words**	**PHQ-9**	-	-	-
14.		**CES-D**	*.99*	-	-
15.	**worry write words**	**GAD-7**	*.58*	*.59*	-
16.		**PSWQ**	*.58*	*.59*	*.97*
			17.	**18.**	**19.**
17.	**depression write phrases**	**PHQ-9**	-	-	-
18.		**CES-D**	*.99*	-	-
19.	**worry write phrases**	**GAD-7**	*.55*	*.56*	-
20.		**PSWQ**	*.55*	*.56*	*.95*
			21.	**22.**	**23.**
21.	**depression write text**	**PHQ-9**	-	-	-
22.		**CES-D**	*.98*	-	-
23.	**worry write text**	**GAD-7**	*.56*	*.56*	-
24.		**PSWQ**	*.51*	*.51*	*.93*

Notes. *N = 963. All correlations are p < .001.*

PHQ-9 = Patient Health Questionnaire-9 assessing depression; CES-D = The Center for Epidemiological Studies Depression Scale (CES-D); GAD-7 = Generalized Anxiety Disorder - 7; PSWQ = Penn State Worry Questionnaire.

Table 5.
Discriminant Validity: 10-Fold Cross-Validated Correlations between Language-Based Assessments of Difference Scores and Observed Difference Scores of Rating Scales

Language responses	Response format	PHQ-9 – GAD-7[1]	CES-D – PSWQ[2]
Depression and worry	**All**	.38***	.43***
Depression		.22***	.37***
Worry		.18***	.24***
Depression and Worry	**Select words**	.37***	.40***
	Write words	.30***	.32***
	Write phrases	.24***	.30***
	Write text	.14***	.36***

Notes. $N = 963$. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.
[1] Predicting the difference score of the normalized PHQ-9 minus the normalized GAD-7, where normalization was achieved by respectively subtracting the column mean from each score and dividing by the column standard deviation.
[2] Predicting the difference score of the normalized CES-D minus the normalized PSWQ, in the same way as the "PHQ-9 - GAD-7". For results based on single responses formats see Table S3.

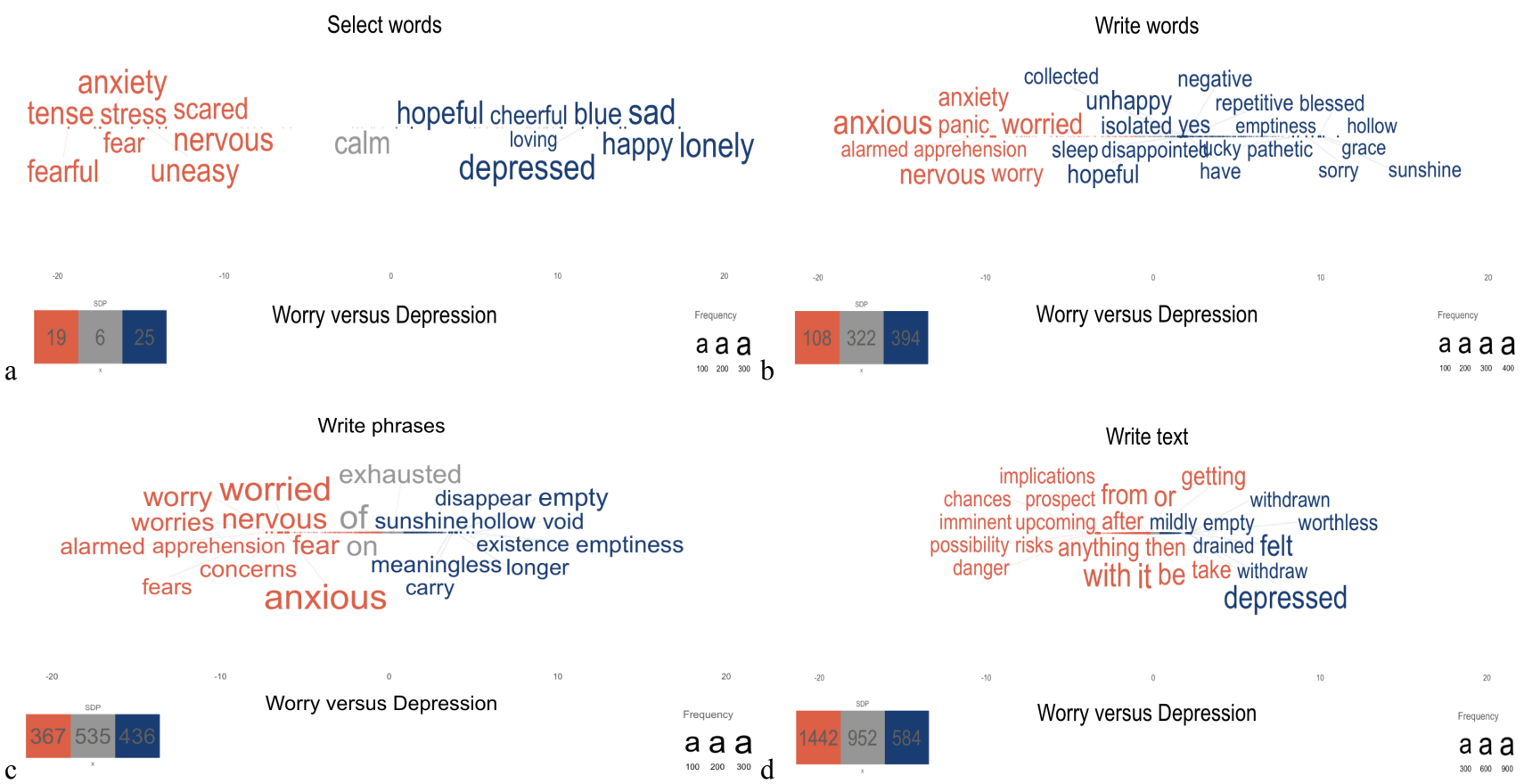


Figure 5a-d | Statistically significant words related to worry (orange) versus depression (blue) responses across the response formats, including a) select words, b) write words, c) write phrases, and d) write text. The more open the word responses are, the more statistically significant words there are. The open-ended response formats include more function words. The open-ended response formats include more function words. Coloured words are statistically significant when correcting for multiple comparisons using “False Discovery Rate”.

Prospective sample reliability: Pre-registered models perform well on new, prospective data

All of the pre-registered models performed above (r = .60-.79)[3] the pre-registered cut-off (r = .50) on the new prospective data (Table 6); in fact, the performance tended to be slightly higher than the cross-validated estimates, where the mean difference between the prospective performance and the cross-validated estimate was .02 (*SD* = .03, *range* = -.03 – .08).

Table 6.
Prospective Sample Reliability: The Pearson Correlation of Single Responses Formats From Pre-Registered Models

	Depression Prompt				Worry Prompt			
	PHQ-9		CES-D		GAD-7		PSWQ	
Response format	Prosp.	CV	Prosp.	CV	Prosp.	CV	Prosp.	CV
Select words	.72	.73	.79	.77	.75	.67	.72	.66
Write words	.69	.68	.76	.73	.65	.67	.64	.66
Write phrases	.72	.69	.76	.75	.66	.62	.67	.61
Write text	.66	.69	.73	.74	.63	.59	.60	.59

Notes. *N = 145.* CV correlations are from Table 3.
PHQ-9 = Patient Health Questionnaire-9 assessing depression; CES-D = The Center for Epidemiological Studies Depression Scale (CES-D); GAD-7 = Generalized Anxiety Disorder - 7; PSWQ = Penn State Worry Questionnaire. For results where depression responses are trained to anxiety and worry scales and vice versa see Table S9.

Test-retest reliability

The test-retest reliability over two weeks was moderate to strong for the pre-registered models in the prospective dataset (Table 7). The correlation for assessing depression from depression prompts with single formats was r = .66 – .75, and for assessing worry/anxiety from worry prompts was r = .52 – .62, whereas the reliability for the rating scales was all very strong (r = .85 – .89).

[3] The registered models for PSWQ were first based on reversing the items incorrectly by subtracting 6 rather than 4 from the item score; we have retrained the models using the correct total scores. For transparency, the incorrect models are still open and called *model_name*_incorrect, and the newly updated models updated after the first analysis are called *model_name*_corrected. In the main document, we are reporting the results for the correct models (though note that it did not make a big difference).

Table 7.
Test-Retest Reliability of the Pre-registered Models: The Pearson Correlation between Time 1 and 2

Response format	Depression Prompt		Worry Prompt	
	PHQ-9	CES-D	GAD-7	PSWQ
Select words	.69	.69	.60	.55
Write words	.74	.75	.52	.54
Write phrases	.71	.72	.62	.60
Write text	.66	.66	.53	.52
All 4[1]	.75	.76	.69	.66
All 8[1]	.76	.77	.73	.67
Rating scale	.85	.87	.86	.89

Notes. N = 122.
PHQ-9 = Patient Health Questionnaire-9 assessing depression; CES-D = The Center for Epidemiological Studies Depression Scale (CES-D); GAD-7 = Generalized Anxiety Disorder-7; PSWQ = Penn State Worry Questionnaire
[1] = These models were not pre-registered. For results where depression responses assess anxiety and worry scales and vice versa see Table S10.

External validity in the prospective sample: Pre-registered model assessments correlate with self-reported sick-leave and healthcare visits

Most response formats yield assessment scores that are weakly to moderately correlated with clinically relevant external criteria related to mental health, including sick leave over the last three months (r = .18 – .28) and the last year (r = .17 – .29), as well as healthcare visits over the last year (r = .18 – .34; Table 8). There is a similar pattern for sick-leave and healthcare visits not specifically related to mental health issues (Table S4). The write text format tended to yield stronger correlations than select words, write words, and write phrases in all but one instance (although the differences tended to be non-significant); this is even though these latter formats tend to be more accurate in predicting rating scales than the write text response format. Further, language-based predictions of the rating scales based on the write text format tend to show slightly higher correlations with the external criteria than the actual observed scores (e.g., self-reported sick leave due to mental health over the last year yields a correlation of r = .26 with the language-based assessments of CES-D, and only a correlation of r = .23 with the observed CES-D). In short, among the language-based response formats, the write text format tends to yield the strongest (or on par with the strongest) correlation to the mental health-related criteria (22 of 24 cases; Table 8) and the general health-related criteria (16 of 24, Table S4). Compared to the observed rating scales, the language-based assessments exhibit an equal or stronger correlation to the externally related criteria of mental health in 9 of 12 cases (Table 8) and for general health in 11 of 12 cases (Table S4).

Table 8.

External Validity in the Prospective Sample: The Pearson Correlation of Single Responses Formats Analysed Using Pre-Registered Models Correlate with Self-Reported External Criteria

	Sick-leave due to mental health issues over the last 3 months				Sick-leave due to mental health issues over the last year				Healthcare visits due to mental health issues over the last year			
Response format	**Language-based assessments**				**Language-based assessments**				**Language-based assessments**			
Depression Prompt	**PHQ-9^{LBA}**	**CES-D^{LBA}**	**GAD-7^{LBA}**	**PSWQ^{LBA}**	**PHQ-9^{LBA}**	**CES-D^{LBA}**	**GAD-7^{LBA}**	**PSWQ^{LBA}**	**PHQ-9^{LBA}**	**CES-D^{LBA}**	**GAD-7^{LBA}**	**PSWQ^{LBA}**
Select words	.24**	.24**	.22**	.23**	.21*	.22**	.20*	.21*	.27**	.26**	.24**	.23**
Write words	.20*	.20*	.20*	.21*	.17*	.17*	.19*	.19*	.20*	.20*	.19*	.18*
Write phrases	.23**	.23**	.21*	.18*	.14	.14	.14	.12	.26**	.26**	.22**	.20*
Write text	.24**	.28***	.28***	.24**	.22**	.26**	.29***	.25**	.32***	.34***	.32***	.29***
Worry Prompt												
Select words	.12	.12	.11	.09	.13	.10	.12	.11	.12	.12	.11	.09
Write words	.11	.11	.10	.09	.15	.14	.14	.11	.11	.12	.09	.09
Write phrases	.01	.02	.02	-.01	.08	.10	.11	.07	.01	.00	-.02	-.05
Write text	.18*	.19*	.18*	.12	.17*	.17*	.17*	.11	.28***	.30***	.26**	.20*
All 8 (not pre-reg.)	.24**	.26**	.19*	.12	.21*	.22*	.21*	.14	.30***	.31***	.21**	.14
	Rating Scales				**Rating Scales**				**Rating Scales**			
	PHQ-9	**CES-D**	**GAD-7**	**PSWQ**	**PHQ-9**	**CES-D**	**GAD-7**	**PSWQ**	**PHQ-9**	**CES-D**	**GAD-7**	**PSWQ**
	.29***	.30***	.21**	.15	.22**	.23**	.19*	.16	.35***	.31***	.19*	.11

Notes. *N = 145.*

PHQ-9 = Patient Health Questionnaire-9 assessing depression; CES-D = The Center for Epidemiological Studies Depression Scale (CES-D); GAD-7 = Generalized Anxiety Disorder - 7; PSWQ = Penn State Worry Questionnaire. not pre-reg. = models not being pre-registered.

*** = $p < .001$, ** = $p < .01$, * = $p < .05$.

Information content across response formats

The response formats differ in the information content they capture (Table 9). The write phrases format yields the highest information content (Diversity Index = 561.0 for depression and 532.2 for anxiety), whereas the select words format has the lowest (Diversity Index = 23.7 for depression and 26.4 for anxiety). It is noteworthy that the information content of rating scale responses is smaller than that from the responses of open-ended formats (the write words, texts, and phrases formats) but larger than the select words format.

Table 9.
Information Content: The Diversity Index of Responses from the Different Assessment Formats

Response Format	Diversity index[1]			
	Depression		Worry	
	Dev.	Prosp.	Dev.	Prosp.
Select words	23.7	27.2	26.4	27.3
Write words	389.4	215.7	343.1	189.0
Write phrases	561.0	312.7	532.2	337.5
Write text	437.3	326.9	460.9	342.5
Rating scales	**Depression**		**Worry/Anxiety**	
PHQ-9	60.7	59.9	-	-
CES-D	45.7	46.2	-	-
GAD-7	-	-	47.1	46.7
PSWQ	-	-	51.6	52.1

Notes. N = 963 for the Development (Dev.) set; N = 145 for the Prospective (Prosp.) set.
[1] the power of $2^{\text{Shannon entropy}}$ is used (see O.N.E. Kjell et al., 2024)

Time burden: The time taken to complete each response format

The survey design enabled us to measure the response time describing both constructs for each response format. Response times with a Z-score of ±3.29 were set to the value corresponding to that Z-score[4]. The median completion time of the select words format for both depression and worry (Median = 61 sec.) was significantly smaller than the aggregated completion time of the PHQ-9 and the GAD-7 (Median = 85 sec., t = -14.05, p < .001) and the CES-D and PSWQ (Median = 135 sec., t = -36.59, p < .001). The second to fastest natural language response format was the write words (Median = 134 sec.), followed by the write phrases (Median = 159 sec.), and last, the write text format (Median = 243 sec.). The response time for the response formats varies widely, where the write text format, on average, takes more than four times as long as the select words format.

[4] Not adjusting for outliers still yields a significant difference: The median completion time of the select words format for both depression and worry (Median = 61 sec.) was significantly smaller than the aggregated completion time of the PHQ-9 and the GAD-7 (Median = 85 sec., t = -6.78, p < .001) and the CES-D and PSWQ (Median = 135 sec., t = -19.07, p < .001).

Table 10
Time Burden: Completion Time (in Seconds) to Answer the Different Response *Formats*

Measure		Mean*		SD		Median	
		Dev.	**Prosp.**	**Dev.**	**Prosp.**	**Dev.**	**Prosp.**
Depression + worry	**Select words**	75	62	46	30	61	55
	Write words	170	148	117	91	134	120
	Write phrases	213	167	159	104	159	133
	Write text	307	217	209	121	243	180
PHQ-9 + GAD-7		100	78	54	40	85	66
CES-D + PSWQ		156	138	77	62	135	120
PHQ-9		55	41	31	20	47	37
CES-D		79	67	40	28	68	61
GAD-7		42	35	25	21	35	29
PSWQ		74	67	40	32	64	58

Notes. *N* = 963 for the Development (Dev.) set; *N* = 145 for the Prospective (Prosp.) set.
PHQ-9: Patient Health Questionnaire-9; CES-D: The Center for Epidemiological Studies Depression Scale; GAD-7: Generalized Anxiety Disorder - 7; PSWQ: Penn State Worry Questionnaire. Numbers are in seconds for “median” and “mean”.
* The survey program did not record the completion time of depression formats and anxiety formats separately.

Table 11.
Overview summary comparing the best across response formats

Response formats	*r mean (range)*						Median time (sec.)[4]	Diversity index[5]
	Predicting rating scales[1]	Discriminant validity[2]	LBA: depression-anxiety[3]	Prospective sample reliability	Test-retest reliability	External criteria		
all 8	.78 (.74-.83)	.90 (.85 - .95)	.41 (.38 - .43)	.78 (.72 - .85)[9]	.73 (.67 - . 77)[10]	.21 (.12 - .31)	488	646
all 4	.76 (.71-.81)	.66 (.64 - .68)	.25 (.18 - .37)	.77 (.74 - . 83)	.71 (.66 - .76)	-	244	571
2 formats[6]	.73 (.65 - 81)	.63 (.61 - .64)	.21 (.11 - .40)	-	-	-	122	434
2 constructs[7]	.71 (.65-.79)	.92 (.88 -.95)	.30 (.14 - .40)	-	-	-	122	208
Single formats[8]								
select words	.66 (.56-.77)	.63 (.62 - .65)	.23 (.17 - .35)	.69 (.55 - .79)	.62 (.55 - .69)	.17 (.09 - .27)	28[11]	24
write words	.64 (.55-.73)	.58 (.58 - .59)	.18 (.14 - .23)	.65 (.57 - .76)	.61 (.45 - .75)	.15 (.09 - .21)	60	389
write phrases	.62 (.51-.75)	.56 (.55 - .56)	.15 (.09 - .23)	.66 (.59 - .76)	.66 (.60 - .72)	.11 (-.05 - .26)	67	561
write text	.61 (.50-.74)	.54 (.51 - .56)	.18 (.07 - .32)	.61 (.48 - .73)	.57 (.52 - .66)	.24 (.11 - .34)	90	437

Notes.
[1] The mean and range is based on correlations based on depression language predicting depression rating scales; and worry language predicting worry/anxiety rating scales (i.e., dark black numbers in Table 3).[2] The correlations between language-based assessments of depression and worry come from Table 4 & Table S2. Correlations of 2 formats include language from two responses of different constructs. [3] Averaged correlations from Table 5 & Table S3. [4] Averaged median completion time from Table 10 on the prospective sample. [5] The diversity indexes of single formats are from the average of development data in table 9. [6] The mean of all two response format combinations from Table S11. [7] The mean of two constructs with the same response formats from Table S11. [8] The mean across response formats predicting corresponding rating scale from Table 3 & Table S8. [9] The means are from Table S9. [10] The means are from Table S10. [11] Since we only recorded the time for answering depression and worry together, we have divided it with here 2 for single response formats.

Discussion

The present study evaluated several validity and reliability aspects of language-based response formats ranging from more closed-ended to open-ended, using the *Sequential Evaluation with Model Pre-registration* (SEMP) design. Overall, the language-based assessments yielded high validity against self-reported rating scales and self-reported external criteria, and moderate to strong reliability, where the different response formats come with different strengths and weaknesses. For example, in the prospective sample, the pre-registered models demonstrated high concurrent validity with depression and worry/anxiety rating scales, which all performed above the pre-registered cut-off. The select words format tended to yield the highest assessment accuracy to rating scales, closely followed by the other formats. These findings align with previous research on the utility of natural language in capturing subjective experiences (e.g., Kjell et al., 2024; K. Kjell et al., 2021), emphasizing its potential to complement traditional methods assessing depression/worry in clinical settings. The results demonstrate that both simple, select-based formats and more elaborate, open-ended text responses can achieve high levels of accuracy, highlighting their viability for diverse clinical and research applications, ranging from rapid assessments in time-constrained settings to in-depth evaluations for diagnostic purposes.

Further, the open-ended response formats showed incremental validity. When combined, they yielded predictive accuracies of rating scales, often approaching the rating scales' own reliability. Combining the eight response formats typically exhibited superior performance, followed by four formats, two formats, and lastly, one response format. These findings highlight the value of integrating diverse response types to achieve a more robust assessment of these mental health constructs compared to relying on a single format.

Challenges and possibilities in discriminating constructs

Language-based assessments based on responses from *different* construct questions yield lower correlations between depression and anxiety than rating scales (language-based assessments: r = .51-.65, versus rating scales: .64-.85). However, using the *same* responses (i.e., from the same construct question(s)) for assessing the two different constructs produces scores that correlate very to extremely strongly (mean r = .92, range .88 – .95), indicating poor discriminant validity between depression and anxiety. Hence, although the models with the same language responses as input produce the highest assessment accuracy, it appears to be at the expense of discriminant validity. This is likely due to the shared variance in the data, where using the same language to assess a construct can be compared with attempting to assess two different constructs using the same item responses from one rating scale. Hence, future language-based assessment models should minimise shared variance by assessing distinct constructs through separate questions and responses (as is done in current practices with different rating scales), while exercising caution when using the same language responses for assessing multiple constructs.

Interestingly, though, it is possible to develop models that differentiate constructs by assessing the standardised difference scores of the rating scales. Hence, research could explore whether these discriminative models can provide insights that enhance treatment planning, such as enabling the offering of more precise and tailored interventions. Improved differentiation has the potential to enable clinicians to better tailor interventions to target the underlying problems. Distinguishing between depression and anxiety is particularly important given their overlapping

symptoms and co-occurrence. By identifying distinct language markers for each condition, as highlighted by Stade et al. (2023), models can improve diagnostic accuracy and guide targeted therapeutic approaches. Furthermore, distinguishing between constructs can enhance symptom monitoring in longitudinal tracking, potentially allowing for a more precise understanding of symptom changes and ensuring timely and appropriate adjustments to treatment plans.

Converging with external criteria

The pre-registered models showed external validity to self-reported sick-leave and healthcare visits. A large body of research shows that depression and anxiety are related to sick leave (for a systematic reviews see Amiri et al., 2021; for a longitudinal study see Sandi et al., 2021 and for a clinical trial see Bjørkedal et al., 2023) and healthcare visits (Cicek et al., 2022; Keller et al., 2018) providing support for the external validity of our measures. Interestingly, the write text format tended to yield the strongest correlations, which were in 9 of 12 cases actually slightly higher or equal to the correlation of observed rating scale scores. Notably, though, all three external validity measures were skewed with a median response of 0 in the current sample. This zero inflation could impact the interpretation of relationships, potentially masking the correlational strength, particularly in the non-zero portion of the sample.

Face validity and the interpretive power of open-ended responses

The language-based assessments can describe respondents' answers with statistically significant words across dimensions such as low to high rating scale scores. At the individual response level, the face validity of open-ended language can reveal whether a respondent has taken the task seriously (e.g., not entering random characters to fill a text box) and whether they have understood the questions (e.g., using “Yes” or “No” instead of providing descriptive answers). This capacity for nuanced evaluation is particularly advantageous compared to closed-ended formats, where insincere responses or misinterpretations of questions or instructions may be much harder to detect. The potential of evaluating the face validity of language responses could become an important aspect in clinical practice, where it can be more straightforward to identify cases in which an individual has not answered the assessment sincerely and diligently, thereby enabling timely intervention or clarification. These insights are harder to detect in closed-ended responses (although one can look for unusual response patterns).

At the group level, visualising open-ended language responses provides a unique source for examining the face and content validity of responses. For example, in all four response formats, respondents described their depression with words such as *blue*, *depressed*, and *emptiness*, and their anxiety with words such as *nervous*, and *panic* – which is according to the DSM-V criteria (“depressed mood”, “diminished interest or pleasure” for major depressive disorder, and “excessive anxiety and worry” and “difficult to control the worry” for generalised anxiety disorder; American Psychiatric Association, 2013). Notably, the more open-ended response formats (i.e., texts, phrases, and words rather than select words) revealed more significant words. Further, the write words and phrases formats tended to contain fewer function words and more content words than the text format. Depending on how important describing individuals’ responses is, these aspects may be important to consider.

Further, as demonstrated in previous research (Kjell et al., 2021) and highlighted in our supplementary material, open-ended responses of depression and anxiety frequently include references to “pain,” “pains,” and “painful” (Table S13), which tend to be significantly related to

higher rating scale scores. This, further underscore the potential of language-based assessments to capture aspects of mental health beyond closed-ended rating scales. Hence, the richer insights provided by more open-ended formats (e.g., texts, phrases, and words) highlight their ability to more comprehensively capture symptoms, which can be valuable in research and clinical settings.

Test-retest reliability and sensitivity to change

The pre-registered models demonstrated moderate to strong test-retest reliability across two weeks, whereas the rating scales' test-retest reliability was very strong. This difference highlights a trade-off between the very strong consistency of closed-ended methods and the higher flexibility of language-based approaches. The lower reliability of language-based assessments might be due to several reasons: First, open-ended language responses have a larger range, openness, and dimensionality than closed-ended rating scale responses (O.N.E. Kjell et al., 2024), which enables them to vary more. Second, the very strong test-retest reliability of rating scales might reflect a lack of sensitivity to change, whereas the moderate to strong test-retest of language-based assessments potentially reflects actual changes in depression and anxiety over two weeks; in other words, it could be that language-based assessments may be more equipped to capture changes. While high test-retest reliability is generally desirable as it indicates consistency, there is a point at which it can signal that a mental health assessment is not sensitive enough to detect changes in a person's condition, lacks nuance in capturing the complexity of symptoms, or both. Detecting changes in symptoms is particularly important in both research and clinical practice, especially during the course of therapy, as it informs treatment effectiveness and necessary adjustments. Future research should explore the utility of language-based assessments for repeated, longitudinal evaluations in clinical settings to determine their sensitivity and practicality in capturing meaningful symptom changes over time.

Optimizing sample sizes for model accuracy

Understanding how model accuracy evolves with increasing training data size helps inform decisions about the minimum number of participants needed to develop robust assessments, while also showing whether current models perform as well as possible given the training sample size. Our results show clear differences in the data requirements for achieving good performance across response formats. This information can optimise resource allocation while ensuring the validity of assessment outcomes. Our results show that models based on the select words format achieve high accuracy with as few as 100 participants, while the writing-based formats require more examples to reach stable performance. As the number of participants increases beyond 500 to 700, all models begin to stabilize, with minimal further improvement in accuracy. This flattening suggests that increasing the sample size beyond this point would not substantially enhance model performance. This plateau indicates that some response formats, such as the select words format, are more efficient for training models on smaller datasets, making them particularly suitable for studies with limited resources. This insight can guide future resource allocation, allowing researchers to prioritize formats that achieve robust performance with fewer participants.

Balancing response time and information content

The response formats vary in response time. More open response formats require more time whilst having the highest ecological validity (i.e., being how we normally communicate complex psychological experiences). This trade-off between brevity and richness is an important

consideration for tailoring assessments to specific contexts. The use of brief assessments is increasingly favored in various research settings; for example, in extensive online surveys where participants may lack the endurance for lengthy assessments, in longitudinal studies involving repeated measures over time, and in initial screenings aimed at rapidly identifying specific characteristics or conditions prior to admitting participants into a comprehensive study (e.g., see Sandy et al., 2017).

The diversity index could explain the increased accuracy of combining response formats in assessment. In general, combining response formats yields higher self-information (a higher diversity index) – and combining more response formats yields higher assessment accuracy. However, whereas the select words format yields the strongest assessment accuracy, it has the lowest diversity index. This might be explained by how the words were selected to be the most frequent answers from a previous study (O.N.E. Kjell et al., 2019). So, the information seems to have been optimized to capture the variance in the rating scale – but it might not be as good in other related outcomes, such as what is seen in their correlation to external criteria, including self-reported sick-leave and health visits; this requires further research.

Clinical relevance and applications in practice

Considering that language comprises higher range, resolution, dimensionality, openness and information content than rating scales (O.N.E. Kjell et al., 2024), it has the potential to be more accurate in converging with important clinically relevant behaviours (i.e. external validity), and thus capturing more variance with signs and symptoms. Effectively getting this information extracted with standardised tools is still an ongoing effort, and particularly few works have explored different natural language response formats.

The results from our study indicate that language-based assessments can provide accurate severity scores for depression and anxiety as well as clinically meaningful descriptions. The severity scores from language-based assessments moderately to strongly converge with rating scales, and the *write text format* showed similar or even higher external validity compared to self-reports for the criterion variables. However, the write text format is also the most time-intensive. Its use may therefore be more suitable in contexts where a comprehensive understanding of the patient's mental health is essential, such as during intake or diagnostic evaluations. Conversely, faster formats like select or write descriptive words are more practical in time-constrained settings or when frequent repeated assessments are required.

Beyond the severity scores, the open-ended response formats also comprise an individual's unique descriptions of symptoms (e.g., *I feel overwhelmed*) and experiences (e.g., *I'm avoiding talking to others because my emotions are unpredictable and I worry that the mask might slip*). These descriptions can be presented to clinicians to support them in reaching a thorough understanding of a patient's condition during, for example, screening, intake or follow-up. The descriptions can also help clinicians identify specific areas or topics to explore in therapy, such as addressing expressed emotions, or concerns. By providing a detailed and personalized picture of a patient's mental state, these insights can guide the planning of therapy sessions, ensuring that the topics addressed are tailored to the patient's unique experiences and needs. Hence, language-based assessments can support person-centered care by providing comprehensive, person-centered insights into their mental health, aligning with the aim of prioritizing individuals' values, needs and preferences in care planning (American Geriatrics Society Expert

Panel, 2015). Additionally, in the course of treatment, these descriptions can be shared with patients to help them reflect on and better understand the trajectory of their mental health (Folkersma et al., 2021; Morris et al., 2010). The potential therapeutic utility of this approach warrants further research.

The different characteristics of the response formats further increase the clinical relevance of language-based assessments by offering different clinical applications. In clinical situations with time constraints, clinicians might prioritise faster response formats, such as the *select* or *write* descriptive words formats, where respondents complete the former faster than the corresponding rating scales. In situations where accuracy and a thorough understanding of the patients are crucial, combining response formats, such as using the *select words* and the *write text* formats. The *select words* format offers a quick and accurate option, and combining it with the *write text* format provides incremental validity and a deeper, more personal assessment, with higher chance of uncovering unique complex experiences. Further, when differentiating between mental illnesses is essential, it is important to use different language questions in the assessments (like one uses different rating scales to assess depression versus anxiety) or use a model trained to yield a difference score.

Language-based assessments support and align with the digital transformation of healthcare (Warraich et al., 2018). Patients could, for example, be invited online to complete a survey where they openly describe their mental health before a meeting, which will not take time from the clinician's work, and it can allow the clinician to prepare the session by reviewing both a severity score and a personal description of the patient's unique experiences and symptoms. Additionally, language-based assessment can be clinically valuable in app-based interventions (Bakker et al., 2016), where patients gain insights into their longitudinal time. Such insights may empower patients to better understand their unique experiences and mental health patterns and progress, fostering engagement and self-management. Further, the time spent on describing one's mental health can potentially be seen as a type of *expressive writing* intervention in itself, leading to improved health (e.g., Lepore, 1997; Lepore et al., 2002; Pennebaker & Beall., 1986; Smyth, 1998); however, to what extent this is the case for our specific open-ended measures requires future research. Future studies could investigate whether the therapeutic effects of expressive writing extend to these open-ended formats, enhancing their dual value as both assessment tools and interventions. Future research could also aim to disentangle this potential effect when evaluating therapeutic interventions, as it is important to distinguish the benefits derived specifically from the intervention itself versus those potentially arising from the act of completing the open-ended assessment.

Lastly, evaluating and visualising the language patterns driving language-based assessment models—such as through word figures—can provide valuable insights into how AI interprets and assesses depression or anxiety severity from language. These insights can be particularly valuable in clinical settings, such as diagnostic evaluations, where understanding the linguistic markers of mental health can support clinicians in identifying key symptoms. Additionally, in research contexts, these visualisations can help elucidate the psychological mechanisms underlying these disorders, providing a deeper understanding of the relationship between language use and mental health.

Limitations and Directions for Future Research

Language-based assessments, like rating scales, are not objective truths of psychological constructs. Like research developing stronger rating scales, our work evaluates approaches to using language-based assessments that meet more criteria of validity and reliability to get closer to latent true scores. However, more extensive evaluations are available and should be done to further establish their validity and reliability such as their ability to converge with best-estimate assessments (e.g., Eijsbroek et al., 2024; Spitzer, 1983) as well as construct relevant observable (i.e., not only self-reported) behaviours.

Our results focus on overall accuracy metrics rather than metrics for individual preferences toward specific response formats. There also could be systematic *individual differences*, where certain individuals may express themselves better with certain response formats – or situational factors influencing the accuracy of the response formats differently. Future research could, for example, examine whether accuracy errors are related to personality traits, educational level, language skill levels, and so on and whether accuracy errors differ across different settings/contexts. Our work lays a foundation informing which response formats seem to work best on average and thus could be built on for work pursuing personalized response formats

The study also has methodological limitations. The questions enabling open-ended responses were presented before the list of words and closed-ended rating scales. We implemented this order to avoid the predefined word lists and items to influence respondents' open-ended answers. However, one possibility is that the open-ended language responses may have introduced priming effects (e.g., Mohan et al., 2016) influencing how they answer the rating scale items. Hence, this priming effect may overestimate the correlation estimate. Future studies should consider randomising the order of open-ended formats and rating scales to disentangle the effects of question order.

Moreover, since this study includes only online samples, caution is advised when generalising the findings beyond this population and context. Future studies should include diverse clinical populations to enhance the generalizability and applicability of language-based assessments in clinical settings. For instance, incorporating participants from outpatient and inpatient clinics, as well as individuals from different cultural and socioeconomic backgrounds, could provide deeper insights into the robustness and validity of these assessments. Additionally, longitudinal studies in clinical contexts could help establish their utility for monitoring treatment progress and outcomes. Further, language-based assessments need to be validated in specific populations and contexts as well as follow relevant regulations (e.g., see the AI act [Hauglid & Mahler, 2023, and Veale & Zuiderveen Borgesius, 2021], and the CE-mark in the EU [European Commission, 2023]).

Last, the current large language model was not explicitly trained for mental health assessment tasks; future research could examine whether finetuning the models may increase the assessment accuracy even further. However, fine-tuning large language models typically requires substantial amounts of data. Preliminary results from our merged dataset, encompassing over 15,000 probed language responses from multiple studies, have not yielded significant gains in accuracy. These findings, combined with the high accuracies achieved in the current study, suggest that RoBERTa-large demonstrates strong generalization capabilities to mental health language without fine-tuning it. Future research could also explore whether greater accuracy can be achieved using other large language models and different predictive model frameworks. Box 1

demonstrates R-code, exemplifying how the registered language models can be applied in future research using the *text* package (O.N.E. Kjell et al., 2023; see Supplement Table S6 for a list of pre-registered models available for automatic download).

Box 1. Example code for using the open pre-registered models

```
# Install the text package
install.packages("text")
library(text)
textrpp_install()
textrpp_initialize()

# Example text to access
text_to_assess ← "Most of the time, I have a hard time finding
meaning in life. I feel down and blue all the time."

# This function automatically downloads the pre-registered models,
transforms the text into word embeddings, and applies the model to
the word embeddings
prediction ← textAssess(
    texts = text_to_assess,
    dim_name = FALSE,
    model_info ←
"depression_text_phq9_roberta23_gu2024")
```

Conclusions

The results provide strong evidence demonstrating concurrent, incremental, discriminant, face, and external validity of the different response formats. Using the *Sequential Evaluation and Model Pre-registration* design, we find that the pre-registered models produce robust prospective sample reliability and test-retest reliability. We show how the response formats differ in, for example, accuracy, visualizations, time burden, and information content. The select response format exhibited superior performance across a range of metrics, including assessment accuracy and response time – however, a list of words is not the natural way of communicating complex psychological phenomena; it produces constrained visualizations and less external validity than text responses. If high accuracy matters, combining the select words and write text formats could be an option. The overall high validity and reliability across the response formats provide the possibility to choose formats according to different research and clinical needs varying in accuracy requirements, time constraints, and response openness.

Transparency

Author Contributions:

Zhuojun Gu: Conceptualization; Methodology; Software; Formal analysis; Data curation; Investigation; Resources; Visualization; Writing – original draft; Writing – review & editing.

Katarina Kjell: Conceptualization; Methodology; Writing – review & editing.

H. Andrew Schwartz: Conceptualization; Methodology; Software; Writing – review & editing.

Oscar Kjell: Conceptualization; Project administration; Funding acquisition; Methodology; Resources; Software; Writing – review & editing.

ORCID iDs

Zhuojun Gu https://orcid.org/0009-0000-1610-4830

Katarina Kjell https://orcid.org/0000-0001-6744-3592

H. Andrew Schwartz https://orcid.org/0000-0002-6383-3339

Oscar Kjell https://orcid.org/0000-0002-2728-6278

References

American Psychiatric Association, D. S. M. T. F., & American Psychiatric Association, D. S. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (Vol. 5, No. 5). Washington, DC: American psychiatric association.

Amiri, S., & Behnezhad, S. (2021). Depression symptoms and risk of sick leave: a systematic review and meta-analysis. *International archives of occupational and environmental health*, *94*(7), 1495-1512.

Bakker, D., Kazantzis, N., Rickwood, D., & Rickard, N. (2016). Mental health smartphone apps: review and evidence-based recommendations for future developments. *JMIR mental health*, *3*(1), e4984.

Baxter, A. J., Scott, K. M., Vos, T., & Whiteford, H. A. (2013). Global prevalence of anxiety disorders: a systematic review and meta-regression. *Psychological medicine*, *43*(5), 897-910.

Bjørkedal, S. T., Fisker, J., Hellström, L. C., Hoff, A., Poulsen, R. M., Hjorthøj, C., ... & Eplov, L. F. (2023). Predictors of return to work for people on sick leave with depression, anxiety and stress: secondary analysis from a randomized controlled trial. *International Archives of Occupational and Environmental Health*, *96*(5), 715-734.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Boyd, R. L., & Schwartz, H. A. (2021). Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, *40*(1), 21-41.

Lépine, J. P., & Briley, M. (2011). The increasing burden of depression. *Neuropsychiatric disease*

and treatment, *7*(sup1), 3-7.

Chevance, A., Ravaud, P., Tomlinson, A., Le Berre, C., Teufer, B., Touboul, S., ... & Tran, V. T. (2020). Identifying outcomes for depression that matter to patients, informal caregivers, and health-care professionals: qualitative content analysis of a large international online survey. *The Lancet Psychiatry*, *7*(8), 692-702.

Cicek, M., Hayhoe, B., Otis, M., Nicholls, D., Majeed, A., & Greenfield, G. (2022). Depression and unplanned secondary healthcare use in patients with multimorbidity: a systematic review. *Plos one*, *17*(4), e0266605.

Coppersmith, G., Dredze, M., & Harman, C. (2014, June). Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 51-60).

Cruz, M., & Pincus, H. A. (2002). Research on the influence that communication in psychiatric encounters has on treatment. *Psychiatric Services*, *53*(10), 1253-1265.

Cutler, A., & Condon, D. M. (2023). Deep lexical hypothesis: Identifying personality structure in natural language. *Journal of Personality and Social Psychology*, *125*(1), 173.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., ... & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), 688-701.

Folkersma, W., Veerman, V., Ornée, D. A., Oldehinkel, A. J., Alma, M. A., & Bastiaansen, J. A. (2021). Patients' experience of an ecological momentary intervention involving self-monitoring and personalized feedback for depression. *Internet Interventions*, *26*, 100436.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., ... & Seligman, M. E. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological science*, *26*(2), 159-169.

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.

Eijsbroek, V. C., Kjell, K., Schwartz, H. A., Boehnke, J. R., Fried, E. I., Klein, D. N., ... & Kjell, O. N. (2024). The LEADING Guideline. Reporting Standards for Expert Panel, Best-Estimate Diagnosis, and Longitudinal Expert All Data (LEAD) Studies. *medRxiv*, 2024-03.

European Commission. (2023). *CE marking*. https://single-market-economy.ec.europa.eu/single-market/ce-marking_en

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of affective disorders*, *208*, 191-197.

Greenberg, L. S., & Pascual‐Leone, A. (2006). Emotion in psychotherapy: A practice‐friendly research review. *Journal of clinical psychology*, *62*(5), 611-630.

Gumus, M., DeSouza, D. D., Xu, M., Fidalgo, C., Simpson, W., & Robin, J. (2023). Evaluating the utility of daily speech assessments for monitoring depression symptoms. *Digital health*, *9*, 20552076231180523.

Hauglid, M. K., & Mahler, T. (2023). Doctor Chatbot: The EU's Regulatory Prescription for Generative Medical AI. *Oslo Law Review*, (1), 1-23.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annu. Rev. Clin. Psychol.*, *3*(1), 29-51.

Keller, A. O., Hooker, R. S., & Jacobs, E. A. (2018). Visits for depression to physician assistants and nurse practitioners in the USA. *The journal of behavioral health services & research*, *45*, 310-319.

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. N. E., 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*.

Kjell, O. N. E., Daukantaitė, D., Hefferon, K., & Sikström, S. (2016). The harmony in life scale complements the satisfaction with life scale: Expanding the conceptualization of the cognitive component of subjective well-being. *Social Indicators Research*, *126*, 893-919.

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., 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), 3918.

Kjell, O. N. E., Kjell, K., & Schwartz, H. A. (2023). Beyond rating scales: With targeted evaluation, language models are poised for psychological assessment. *Psychiatry Research*, 115667.

Kjell, O. N. E., Ganesan, A. V., Boyd, R., Oltmanns, J. R., Rivero, A., Feltman, S., ... & Schwartz, H. A. Demonstrating High Validity of a New AI-Language Assessment of PTSD: A Sequential Evaluation with Model Pre-registration.

Kwantes, P. J., Derbentseva, N., Lam, Q., Vartanian, O., & Marmurek, H. H. (2016). Assessing the Big Five personality traits with latent semantic analysis. *Personality and Individual Differences*, *102*, 229-233.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Lepore, S. J. (1997). Expressive writing moderates the relation between intrusive thoughts and depressive symptoms. *Journal of personality and social psychology*, *73*(5), 1030.

Lepore, S. J., Greenberg, M. A., Bruno, M., & Smyth, J. M. (2002). Expressive writing and health: Self-regulation of emotion-related experience, physiology, and behavior.

Levinson, S. C., & Holler, J. (2014). The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130302.

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

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (1907). RoBERTa: A robustly optimized BERT pretraining approach. arXiv [Preprint](2019). *arXiv preprint arXiv:1907.11692*.

Lundqvist, C., Jederström, M., Korhonen, L., & Timpka, T. (2022). Nuances in key constructs need attention in research on mental health and psychiatric disorders in sports medicine. BMJ *Open Sport & Exercise Medicine*, *8*(3), e001414.

Lynn, V., Balasubramanian, N., & Schwartz, H. A. (2020, July). Hierarchical modeling for user personality prediction: The role of message-level attention. In *Proceedings of the 58th*

annual meeting of the association for computational linguistics (pp. 5306-5316).
Marengo, D., Azucar, D., Longobardi, C., & Settanni, M. (2021). Mining Facebook data for Quality of Life assessment. *Behaviour & Information Technology*, 40(6), 597-607.
Matero, M., Hung, A., & Schwartz, H. A. (2021). Evaluating contextual embeddings and their extraction layers for depression assessment. *arXiv preprint arXiv:2112.13795*.
Matero, M., Vu, H., Nilsson, A., Mahwish, S., Cho, Y. M., McKay, J., ... & Schwartz, H. A. (2024, March). Using Daily Language to Understand Drinking: Multi-Level Longitudinal Differential Language Analysis. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)* (pp. 133-144).
Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Development and validation of the penn state worry questionnaire. *Behaviour research and therapy*, *28*(6), 487-495.
Mohan, D. M., Kumar, P., Mahmood, F., Wong, K. F., Agrawal, A., Elgendi, M., ... & Chan, A. H. (2016). Effect of subliminal lexical priming on the subjective perception of images: A machine learning approach. *PLoS one*, 11(2), e0148332.
Mokkink, L. B., Prinsen, C., Patrick, D. L., Alonso, J., Bouter, L., De Vet, H. C., ... & Mokkink, L. (2018). COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs). *User manual*, *78*(1), 6-3.
Morris, M. E., Kathawala, Q., Leen, T. K., Gorenstein, E. E., Guilak, F., Labhard, M., & Deleeuw, W. (2010). Mobile therapy: case study evaluations of a cell phone application for emotional self-awareness. *Journal of medical Internet research*, *12*(2), e10.
Palan, S., & Schitter, C. (2018). Prolific. ac—A subject pool for online experiments. *Journal of behavioral and experimental finance*, *17*, 22-27.
Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., ... & Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of personality and social psychology*, *108*(6), 934.
Pennebaker, J. W., & Beall, S. K. (1986). Confronting a traumatic event: toward an understanding of inhibition and disease. *Journal of abnormal psychology*, 95(3), 274.
R Core Team. (2013). R: A language and environment for statistical computing. *Foundation for Statistical Computing, Vienna, Austria*.
Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied psychological measurement*, *1*(3), 385-401.
Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822*.
Rutowski, T., Shriberg, E., Harati, A., Lu, Y., Chlebek, P., & Oliveira, R. (2020, November). Depression and anxiety prediction using deep language models and transfer learning. In *2020 7th International Conference on Behavioural and Social Computing (BESC)* (pp. 1-6). IEEE.
Sandin, K., Anyan, F., Osnes, K., Gjengedal, R. G. H., Leversen, J. S. R., Reme, S. E., & Hjemdal, O. (2021). Sick leave and return to work for patients with anxiety and depression: a longitudinal study of trajectories before, during and after work-focused treatment. *BMJ open*, *11*(9), e046336.
Sandy, C. J., Gosling, S. D., Schwartz, S. H., & Koelkebeck, T. (2017). The development and validation of brief and ultrabrief measures of values. *Journal of personality assessment*, *99*(5), 545-555.

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.

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.

Schwartz, H. A., Eichstaedt, J., Kern, M., Park, G., Sap, M., Stillwell, D., ... & Ungar, L. (2014, June). Towards assessing changes in degree of depression through facebook. In *Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality* (pp. 118-125).

Schwartz, H. A., Sap, M., Kern, M. L., Eichstaedt, J. C., Kapelner, A., Agrawal, M., ... & Ungar, L. H. (2016). Predicting individual well-being through the language of social media. In *Biocomputing 2016: Proceedings of the pacific symposium* (pp. 516-527).

Sennrich, R. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Shannon, Claude Elwood. "A mathematical theory of communication." *The Bell system technical journal* 27.3 (1948): 379-423.

Shen, H., Li, Z., Xu, C., Zhu, J., Chen, M., & Fang, Y. (2018). Analysis of misdiagnosis of bipolar disorder in an outpatient setting. *Shanghai Archives of Psychiatry*, *30*(2), 93.

Smyth, J. M. (1998). Written emotional expression: effect sizes, outcome types, and moderating variables. *Journal of consulting and clinical psychology*, *66*(1), 174.

Spitzer, R. L. (1983). Psychiatric diagnosis: are clinicians still necessary?. *Comprehensive psychiatry*.

Spitzer, R. L., Kroenke, K., Williams, J. B., Patient Health Questionnaire Primary Care Study Group, & Patient Health Questionnaire Primary Care Study Group. (1999). Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *Jama*, *282*(18), 1737-1744.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of internal medicine*, *166*(10), 1092-1097.

Stahnke, B. (2021). A systematic review of misdiagnosis in those with obsessive-compulsive disorder. *Journal of affective disorders reports*, 6, 100231.

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.

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.

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* (Vol. 2021, p. 4515). NIH Public Access.

Vaci, N., Liu, Q., Kormilitzin, A., De Crescenzo, F., Kurtulmus, A., Harvey, J., ... & Nevado-Holgado, A. (2020). Natural language processing for structuring clinical text data on depression using UK-CRIS. *BMJ Ment Health*, *23*(1), 21-26.

Varadarajan, V., Sikström, S., Kjell, O., & Schwartz, H. (2024, June). ALBA: Adaptive

Language-Based Assessments for Mental Health. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 2466-2478).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention Is All You Need.(Nips), 2017. *arXiv preprint arXiv:1706.03762*, *10*, S0140525X16001837.

Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 97-112.

Warstadt, A., Singh, A., & Bowman, S. R. (2019). Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics*, *7*, 625–641.

Warraich, H. J., Califf, R. M., & Krumholz, H. M. (2018). The digital transformation of medicine can revitalize the patient-clinician relationship. *NPJ digital medicine*, *1*(1), 49.

Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. sage publications.