Developing and Validating Language-Based Assessments for Mental Health

Measuring and Describing Depression, Anxiety, Affect, and Suicidality and Self-Harm Risk, from Individuals' Own Descriptions

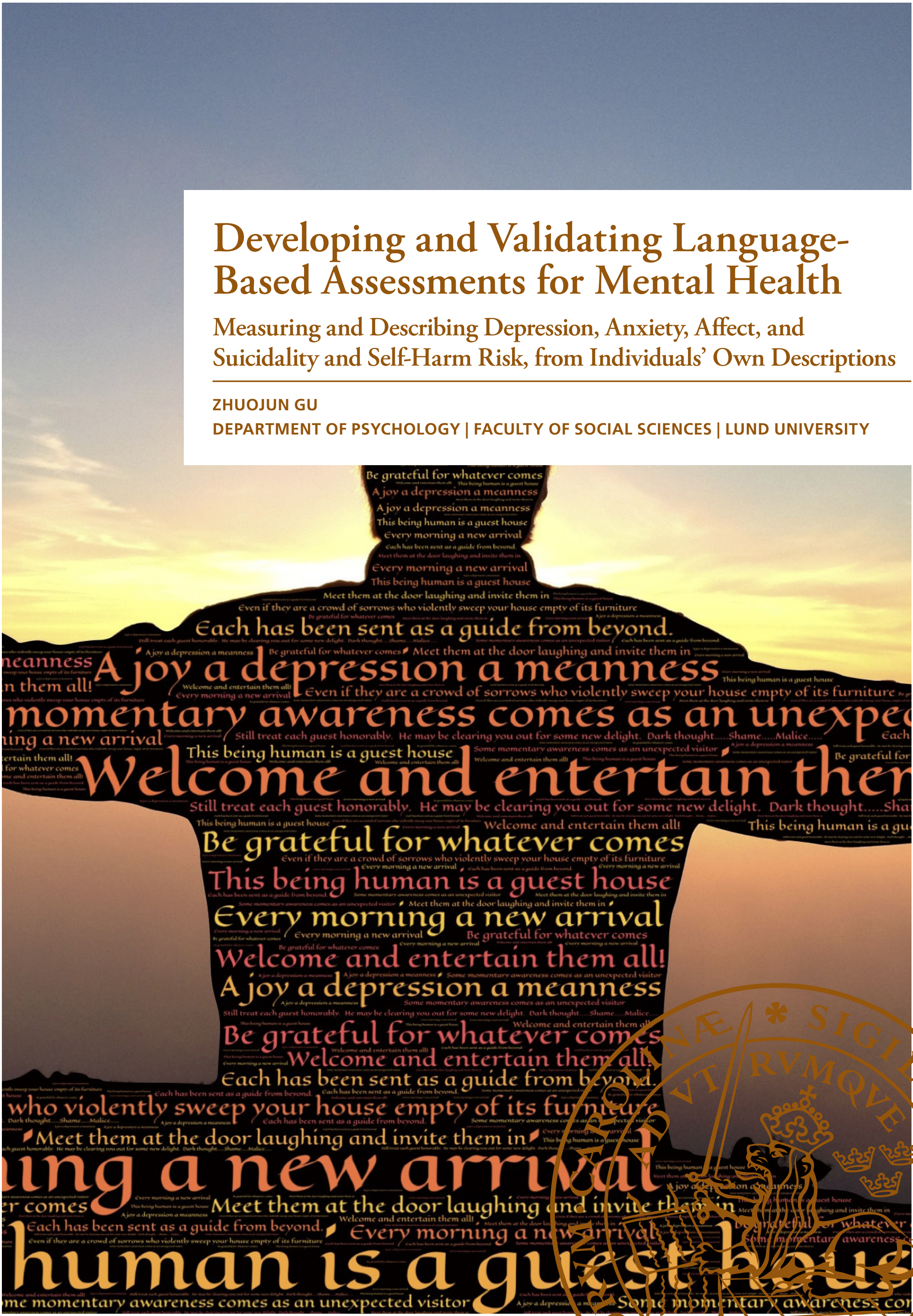
Gu, Zhuojun

2025

Developing and Validating Language-Based Assessments for Mental Health

Measuring and Describing Depression, Anxiety, Affect, and Suicidality and Self-Harm Risk, from Individuals' Own Descriptions

ZHUOJUN GU
DEPARTMENT OF PSYCHOLOGY | FACULTY OF SOCIAL SCIENCES | LUND UNIVERSITY


Be grateful for whatever comes
A joy a depression a meanness
This being human is a guest house
Every morning a new arrival
Each has been sent as a guide from beyond.
Meet them at the door laughing and invite them in
Even if they are a crowd of sorrows who violently sweep your house empty of its furniture
momentary awareness comes as an unexpec
Still treat each guest honorably. He may be clearing you out for some new delight. Dark thought.....Shame.....Malice.....
Welcome and entertain them all!
Some momentary awareness comes as an unexpected visitor
ning a new arrival
human is a guest house

Developing and Validating Language-Based Assessments for Mental Health

Developing and Validating Language-Based Assessments for Mental Health

Measuring and Describing Depression, Anxiety, Affect, and Suicidality and Self-Harm Risk, from Individuals' Own Descriptions

Zhuojun Gu

DOCTORAL DISSERTATION

by due permission of the Faculty of Social Sciences at Lund University, Sweden to be defended on 7th of November 2025 at 13.00 in Sh:128 Gamla Köket, Lund.

Faculty opponent
Prof. Dr. Clements Stachl, University of St. Gallen

Organization:

LUND UNIVERSITY

Author(s): Zhuojun Gu

Document name

Date of issue November 7, 2025

Sponsoring organization

Title and subtitle: Developing and Validating Language-Based Assessments for Mental Health

Measuring and Describing Depression, Anxiety, Affect, and Suicidality and Self-Harm Risk, from Individuals’ Own Descriptions

Abstract:

This thesis develops and evaluates language-based assessments that use artificial intelligence (AI) to transform open-ended language into quantitative indicators and descriptions of mental health related constructs. While closed-ended scales have long dominated psychological assessment, they are limited by fixed response formats and may not fully capture the complexity of individuals’ experiences. By contrast, language offers a flexible, and expressive medium for describing thoughts, emotions, and behaviors. Across four papers, this thesis examines whether language-based assessments can provide valid, and reliable tools for assessing psychological constructs such as depression, anxiety and affect, as well as mental health related risk assessments including suicidality and self-harm.

Paper I compares four different language response formats—from selecting predefined words to producing full-text responses. We evaluated the response formats in terms of their validity—covering concurrent, incremental, face, discriminant, and external aspects—and their reliability, including test-retest and performance in a prospective sample. Using the Sequential Evaluation with Model Pre-registration (SEMP) approach, machine learning models were trained on a development dataset (*N* = 963) and pre-registered before being tested on a separate prospective sample (*N* = 145). These pre-registered models demonstrated moderate to strong validity and reliability, achieving predictive accuracy in the new sample (*r* = .60–.79). The consistent performance across formats suggests that they may be selected based on specific research or potential practical requirements.

Paper II evaluates AI-based language models to evaluate the risk of suicide and self-harm based on individuals’ open-ended narratives about suicidality, self-harm, depression, anxiety, and overall mental health. Employing the SEMP framework, models were trained (*N* = 641) and pre-registered, then validated in a held-out set (*N* = 150) against expert ratings generated using the Longitudinal Expert Data (LED) approach. In a held-out test set, the language-based assessments showed alignment with expert ratings for suicidality (*r* = .70) and self-harm (*r* = .68), and significantly outperformed models that relied on demographic data.

Paper III evaluates the causal validity of language-based assessments in an experimental setting. Few studies have tested whether language-based assessments can detect causal changes. In this randomized mixed-design experiment (*N* = 892), participants underwent mood induction in physical settings (*N* = 153) or via online videos (*N* = 739) across three conditions (church, mall, park). They reported pre- and post-mood affect using both open-ended responses and closed-ended Positive and Negative Affect (PANAS) ratings. We compared how well PANAS and language-based assessments classified the conditions. Language-based assessments outperformed PANAS in predictive accuracy across training (*AUC* = .74 vs. .63), online (*AUC* = .76 vs. .70), and offline holdout samples (*AUC* = .67 vs. .53). In addition, language-based assessments provided qualitative insights by visualizing word-level patterns across conditions.

Paper IV introduces the L-BAM Library—an open repository for pre-validated language-based assessment models—and outlines a framework for sharing and applying these tools in transparent and reproducible ways. This paper emphasizes responsible open-science practices and encourages the independent validation of LBAs in new populations and contexts.

In sum, this thesis demonstrates that language-based assessments can serve as valid, reliable, and informative research tools for measuring and describing mental health-related constructs. By leveraging the expressiveness of open-ended language, these methods address limitations of traditional scales and offer new possibilities for capturing complex psychological phenomena. Through systematic validation across diverse samples and contexts—including expert-rated risk assessment, experimental manipulations, and real-world implementation—this work contributes to the methodological advances of language-based methods into the broader landscape of psychological assessment and highlights the importance of transparent, cumulative practices for their continued development.

Key words: Language-based assessment, Depression, Anxiety, Suicidality, Self-harm, Affect, AI, Open science

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language: English

ISSN and key title

ISBN (print): 978-91-8104-705-9

ISBN (Electronic): 978-91-8104-706-6

Recipient’s notes

Number of pages 104

Price

Security classification

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature

Date 2025-09-23

Developing and Validating Language-Based Assessments for Mental Health

Measuring and Describing Depression, Anxiety, Affect, and Suicidality and Self-Harm Risk, from Individuals' Own Descriptions

Zhuojun Gu

LUND UNIVERSITY

Copyright

Published by:

Department of Psychology

Faculty of Social Sciences

Lund University

Lund 2025

ISBN 978-91-8104-705-9 (print)

ISBN 978-91-8104-706-6 (electronic)

Printed in Sweden by Media-Tryck, Lund University,
Lund, 2025

NORDIC SWAN ECOLABEL
Printed matter
3041 0903

Media-Tryck is a Nordic Swan Ecolabel certified provider of printed material. Read more about our environmental work at www.mediatryck.lu.se

MADE IN SWEDEN

Dedication

To my entire family, and to all my friends. Also, to AI, geometry, mind, psychohistory, and ***the unknown****.*

Table of Contents

List of Abbreviations

AI	Artificial Intelligence
AUC	Area-Under-Curve
BERT	Bidirectional Encoder Representations from Transformers
CES-D	Center for Epidemiologic Studies Depression Scale
DSM	Diagnostic and Statistical Manual of mental disorders
EMA	Ecological Momentary Analysis
GAD-7	Generalized Anxiety Disorder scale-7
GDPR	General Data Protection Regulation
GPT	Generative Pre-trained Transformers
HiTOP	Hierarchical Taxonomy of Psychopathology
LBA	Language-Based Assessment
L-BAM	Language-Based Assessment Model library
LED	Longitudinal Expert Data
LEAD	Longitudinal Expert All Data
LEADING	Longitudinal, Evaluation – experts, materials and procedures, Appropriate Data, and Validity
MIP	Mood Induction Procedure
PANAS	Positive and Negative Affect Schedule
PHQ-9	Patient Health Questionnaire-9
PSWQ	Penn State Worry Questionnaire
RoBERTa	Robustly Optimized BERT Pretraining Approach
SCID	Structured Clinical Interview for DSM Disorders
SEMP	Sequential Evaluation with Model Pre-registration

Acknowledgement

I am deeply grateful for the opportunity to complete this doctoral thesis—a journey filled with intellectual challenges, stimulating collaborations, and the generous support of many people. None of the ideas and work presented here would have been possible without the encouragement, guidance, and companionship of others. For all of this, I feel truly thankful.

First and foremost, I would like to express my sincere gratitude to my supervisors. Oscar Kjell, my main supervisor, has been an invaluable source of insight, rigor, and constructive criticism. Your clarity of thought and unwavering standards has greatly sharpened my work. I also wish to thank Kajsa Järvholm for her thoughtful guidance, encouragement, and attention to detail throughout the process. Finally, I am grateful to my advisor Andrew Schwartz for his inspiration and continuous feedback on both theoretical and methodological aspects. Together, your mentorship has made this work not only possible but deeply rewarding.

I also gratefully acknowledge the support from the Harmony Lab, and the Department of Psychology at Lund University, and the broader community of colleagues and collaborators—both within and outside the university—who provided feedback, engaged in discussions, and helped me refine the studies included in this thesis. Special thanks to my co-authors and collaborators for sharing their expertise and for building projects that combined psychology, artificial intelligence, and open science at the computer science department of Stony Brook University. I would also like to thank the Faculty of Social Sciences and the cross-university project eSSENCE that funded my doctoral studies. I am grateful to the NAISS project for providing the ALVIS computing infrastructure, which was essential for carrying out this work too.

My gratitude extends to the participants who contributed their time and experiences. Their willingness to share their words and perspectives made this research possible and meaningful.

On a more personal note, I am profoundly indebted to my entire family. Thank you for your endless love, support, and belief in me. This thesis is dedicated to you.

Finally, I wish to acknowledge the less tangible but equally important sources of inspiration that have accompanied me throughout this work: artificial intelligence, and language-based assessment. These themes, bridging science and imagination, have fueled my curiosity and sustained my motivation.

To all of you who have been part of this journey—colleagues, friends, and loved ones—thank you.

Abstract

This thesis develops and evaluates language-based assessments that use artificial intelligence (AI) to transform open-ended language into quantitative indicators and descriptions of mental health related constructs. While closed-ended scales have long dominated psychological assessment, they are limited by fixed response formats and may not fully capture the complexity of individuals' experiences. By contrast, language offers a flexible, and expressive medium for describing thoughts, emotions, and behaviors. Across four papers, this thesis examines whether language-based assessments can provide valid, and reliable tools for assessing psychological constructs such as depression, anxiety and affect, as well as mental health related risk assessments including suicidality and self-harm.

Paper I compares four different language response formats—from selecting predefined words to producing full-text responses. We evaluated the response formats in terms of their validity—covering concurrent, incremental, face, discriminant, and external aspects—and their reliability, including test-retest and performance in a prospective sample. Using the Sequential Evaluation with Model Pre-registration (SEMP) approach, machine learning models were trained on a development dataset (*N* = 963) and pre-registered before being tested on a separate prospective sample (*N* = 145). These pre-registered models demonstrated moderate to strong validity and reliability, achieving predictive accuracy in the new sample (*r* = .60–.79). The consistent performance across formats suggests that they may be selected based on specific research or potential practical requirements.

Paper II evaluates AI-based language models to evaluate the risk of suicide and self-harm based on individuals' open-ended narratives about suicidality, self-harm, depression, anxiety, and overall mental health. Employing the SEMP framework, models were trained (*N* = 641) and pre-registered, then validated in a held-out set (*N* = 150) against expert ratings generated using the Longitudinal Expert Data (LED) approach. In a held-out test set, the language-based assessments showed alignment with expert ratings for suicidality (*r* = .70) and self-harm (*r* = .68), and significantly outperformed models that relied on demographic data.

Paper III evaluates the causal validity of language-based assessments in an experimental setting. Few studies have tested whether language-based assessments can detect causal changes. In this randomized mixed-design experiment (*N* = 892), participants underwent mood induction in physical settings (*N* = 153) or via online videos (*N* = 739) across three conditions (church, mall, park). They reported pre- and post-mood affect using both open-ended responses and closed-ended Positive and Negative Affect (PANAS) ratings. We compared how well PANAS and language-based assessments classified the conditions. Language-based assessments outperformed PANAS in predictive accuracy across training (*AUC* = .74 vs. .63), online (*AUC* = .76 vs. .70), and offline holdout samples (*AUC* = .67 vs. .53). In

addition, language-based assessments provided qualitative insights by visualizing word-level patterns across conditions.

Paper IV introduces the L-BAM Library—an open repository for pre-validated language-based assessment models—and outlines a framework for sharing and applying these tools in transparent and reproducible ways. This paper emphasizes responsible open-science practices and encourages the independent validation of LBAs in new populations and contexts.

In sum, this thesis demonstrates that language-based assessments can serve as valid, reliable, and informative research tools for measuring and describing mental health-related constructs. By leveraging the expressiveness of open-ended language, these methods address limitations of traditional scales and offer new possibilities for capturing complex psychological phenomena. Through systematic validation across diverse samples and contexts—including expert-rated risk assessment, experimental manipulations, and real-world implementation—this work contributes to the methodological advances of language-based methods into the broader landscape of psychological assessment and highlights the importance of transparent, cumulative practices for their continued development.

Populärvetenskaplig sammanfattning

Hur vi mår psykiskt påverkar hur vi fungerar i vardagen, men att mäta psykisk ohälsa är inte enkelt. I denna avhandling undersöks hur artificiell intelligens (AI) kan användas för att tolka människors egna ord—så kallat naturligt språk—för att mäta tillstånd som depression, ångest och självmordsrisk. I fyra studier utvecklades och utvärderades AI-modeller som analyserar deltagarnas svar på öppna frågor om deras psykiska mående. Dessa modeller kunde sedan bedömma hur individers mående.

Avhandlingen visar att språkbaserade bedömningar kan vara lika tillförlitliga, eller till och med bättre, än traditionella skattningsskalor—särskilt när det gäller att fånga upp nyanserade eller tillfälliga förändringar i mående. Exempelvis kunde modellerna upptäcka skillnader i känslotillstånd efter att deltagare utsatts för olika miljöer i ett experiment. Arbetet introducerar också nya forskningsmetoder och verktyg, bland annat en öppen bibliotekstjänst kallad L-BAM, där andra forskare fritt kan ta del av modellerna, testa dem i nya sammanhang och bidra till ökad transparens och vetenskaplig reproducerbarhet.

Samtidigt lyfter avhandlingen viktiga begränsningar och etiska aspekter. Språklig förmåga och insikt om det egna måendet varierar mellan individer, vilket kan påverka hur rättvis och träffsäker bedömningen blir. Dessutom är det viktigt att beakta frågor om integritet när AI används för att analysera personliga texter. Sammanfattningsvis visar avhandlingen att AI kan stödja men inte ersätta mänsklig bedömning, och att språk kan vara ett kraftfullt verktyg för att lyssna på människors inre upplevelser—om det används med omtanke.

List of Papers

Paper I

Gu, Z., Kjell, K., Schwartz, H. A., & Kjell, O. (2025). Natural language response formats for assessing depression and worry with large language models: A sequential evaluation with model pre-registration. *Assessment*. Advance online publication. https://doi.org/10.1177/10731911251364022

Copyright © 2025 The Authors. Reprinted by permission of SAGE Publications.

Paper II

Gu, Z., Eijsbroek, V., Kjell, K., Wiebel, C., Järvholm, K., Schwartz, A., & Kjell, O. (unpublished manuscript). Understanding Suicidality and Self-Harm Through Probed Open-Ended Language: A sequential evaluation with model pre-registration.

Paper III

Bång, O., Gu, Z., Nilsson, A. H., Eijsbroek, V., Wiebel, C., Ganesan, A., Kjell, K., Schwartz, H. A., & Kjell, O. (submitted). *Language-based assessments capture experiment-induced emotions more accurately than rating scales*.

Paper IV

Nilsson, A. H., Eijsbroek, V. C., Gu, Z., Kjell, K., Giorgi, S., Kotov, R., Ganesan, A. V., Schwartz, H. A., & Kjell, O. N. E. (under revision). The Language-Based Assessment Model (L-BAM) Library: Open model sharing for independent validation and broader applications. (Resubmitted at *Advances in Methods and Practices in Psychological Science*)

Chapter 1: Background and Objectives

Accurate psychological assessment is a cornerstone of both clinical practice and psychological research. In clinical settings, effective assessments help determine the presence and severity of mental health conditions (Black et al., 2014; Reas et al., 2013), guide treatment decisions (Davis et al., 2018; Feder et al., 2022), monitor therapeutic progress (Pedersen et al., 2013; Yonashiro-Cho et al., 2021), and facilitate communication among professionals and patients (Hunsley, & Mash, 2007). In research contexts, assessments serve as the foundation for testing theoretical models (e.g., Cronbach, & Meehl, 1955), evaluating interventions (e.g., Youngstrom et al., 2015), and identifying risk (e.g., Adams et al., 2024) or protective factors (e.g., Wille et al., 2008) across populations. Many psychological conclusions—whether diagnostic or theoretical—depend in part on the quality of the measurement tools used.

Traditional methods of assessing mental health are widely used but have several well-recognized limitations. Structured and semi-structured clinical interviews—such as the Structured Clinical Interview for Diagnostic and Statistical Manual-5 (The Structured Clinical Interview for DSM-5, SCID, in First, 2014)—are often considered the most accurate individual tools for identifying conditions like depression and anxiety. However, these interviews are resource-intensive, requiring significant time and specialized training, and they are not consistently implemented in everyday clinical practice (Miller et al., 2015; Mueller & Segal, 2015; Navandi et al., under review). Closed-ended rating scales, such as the Patient Health Questionnaire-9 (PHQ-9, Kroenke et al., 2001), are also commonly used by clinicians, but these tools often fall short in capturing the full complexity, depth, and context of an individual's psychological experiences (DeJonckheere, & Vaughn 2019; Kjell et al., 2024; Monson et al., 2016). This disconnects between subjective experience and standardized measurement can lead to concerns about the validity and reliability of traditional assessments (Menold et al., 2018; Schaeffer & Dykema, 2011).

Recent advancements in artificial intelligence (AI) and natural language processing (NLP) offer exploratory opportunities to complement conventional assessments (Bhatia, & Aka, 2022; Boyd, & Schwartz, 2021; De Choudhury et al., 2013; Demszky et al., 2023; Dumas et al., 2025; Eichstaedt et al., 2015; Kjell et al., 2023,

& 2024; Sametoğlu et al., 2024). Language is one of the most direct and expressive ways through which individuals convey their internal states (Boyd, & Schwartz, 2021; Tausczik, & Pennebaker, 2010). These models interpret responses using contextual patterns rather than word frequency alone, offering potential insights into psychological expression (Harrer, 2023; Meskó, & Topol, 2023; Vaswani et al., 2017). While AI-driven models may capture psychological signals from open-ended language, these models are still at an early stage of development and must be interpreted cautiously. Their findings can provide hypotheses and supplement traditional closed-ended measures (He et al., 2023; Kjell et al., 2024), but further validation—especially against behavioral outcomes and clinical interviews—is necessary before considering them for applied clinical use.

Research Objectives and Questions

The overarching aim of this thesis is to develop and evaluate AI-driven language-based assessments for measuring mental health related psychological constructs such as depression, anxiety, and suicidality risk. It seeks to address key challenges in psychological measurement by introducing and evaluating new methods that are not only technically innovative but also psychometrically sound and ethically responsible. Drawing on the theoretical and psychometric frameworks outlined in the following chapters, this work explores whether natural language—when analyzed using modern AI techniques—can offer valid and reliable indicators of psychological states under controlled conditions. More specifically, the thesis addresses the following core objectives:

- **To develop AI-based assessment models** that convert open-ended language responses into quantitative indicators of mental health, using techniques from AI, i.e., natural language processing, large language models and machine learning.
- **To evaluate the psychometric soundness of these models**, focusing on their validity (e.g., construct, criterion, discriminant, face, and causal validity) and reliability (e.g., test–retest, and prospective reliability).
- **To adhere to and further develop open-science practices for language-based assessments** that foster transparency, reusability, and cumulative progress — including pre-registered evaluation frameworks, as well as the creation of tools for sharing models.

Together, these objectives support the broader goal of investigating language-based assessments that are methodologically rigorous, ethically responsible, and potentially scalable for research use. Their use in clinical settings, however, requires further validation across diverse populations and real-world applications.

The structure of the thesis

The structure of this thesis is organized to guide the reader from foundational concepts to applied evaluations. Chapter 2 provides an overview of psychological assessments, covering traditional methods, and key psychometric principles that inform clinical utility. Chapter 3 lays the theoretical foundation for language-based assessments, framing language as both behavior and data, and exploring how advances in AI and natural language processing support this shift. Chapter 4 presents the methodological framework, detailing the models, validation strategies, and experimental designs used across the papers. Chapter 5 summarizes the four papers, each contributing distinct empirical insights into the development, validation, and dissemination of language-based tools. Chapter 6 addresses ethical and regulatory considerations, including issues of privacy, fairness, and clinical transparency in AI-driven assessments. Finally, Chapter 7 concludes the thesis by synthesizing key findings and outlining future directions for research and practice.

Chapter 2: Introduction to Psychological Assessments

Introduction to Psychological Assessments

Psychological assessments are (systematic) processes used to evaluate mental health constructs such as symptoms, behaviors, and functioning, typically for the purposes of diagnosis, treatment planning, and monitoring of clinical change (Black et al., 2014; Youngstrom et al., 2015, & 2017; Reas et al., 2013). In clinical practice, psychological assessments serve multiple core functions: they help measure the severity of psychological symptoms (Bjureberg et al., 2022; Kroenke et al., 2021), differentiate between overlapping diagnostic constructs (DeYoung et al., 2022), inform therapeutic decision-making (Hunsley, & Mash, 2007), and support communication between professionals and patients (Navandi et al., under review; Swets et al., 2000). In research, they are essential for validating theoretical models (Grahek et al., 2021), comparing treatment outcomes (Davis et al., 2018; Feder et al., 2022), and developing new diagnostic tools or interventions (Plake & Wise, 2014).

An accurate psychological assessment is crucial because it directly influences the quality and outcomes of both clinical care and research (Navandi et al., under review; Stefana et al., 2025; Wright et al., 2022). In clinical contexts, accurate assessments enable clinicians to identify the correct diagnosis, determine the severity of a condition, and tailor interventions to the individual's needs—thereby improving treatment efficacy and reducing the risk of misdiagnosis or inappropriate care. Inaccurate assessments, by contrast, can lead to under- or over-treatment, misallocation of healthcare resources, and potentially harmful consequences for individuals (Meyer et al., 2001). In research, accurate measurement ensures that findings about psychological constructs, interventions, or risk factors are valid and generalizable. Without accuracy, conclusions drawn from data may be misleading, compromising both theoretical development and the translation of research into practice (Bossuyt et al., 2003).

Historical and Present Approaches to Mental Health Assessment

Psychological assessment has historically relied on a range of methodologies, broadly categorized into structured, semi-structured, and unstructured clinical interviews, self-report questionnaires and rating scales, and observational and

Unstructured clinical interviews involve open-ended conversations guided by clinician judgment and are flexible but highly subjective (Bihu, 2020). In contrast, structured and semi-structured interviews follow predefined formats that ensure consistency across assessments. These structured formats, such as the Structured Clinical Interview for DSM (First, 2014), offer stronger psychometric properties and are more likely to lead to reliable and valid diagnoses as compared to unstructured clinical interviews (Shankman et al., 2018; Tolin et al., 2018). However, they are often more time-consuming and require formal training to administer.

Self-report questionnaires and rating scales are another cornerstone of traditional psychological assessment. Instruments like the Patient Health Questionnaire-9 (for depression; Kroenke et al., 2001) or the Generalized Anxiety Disorder-7 (GAD-7 for anxiety; Spitzer et al., 2006) are quick to administer, cost-effective, and have well-established reliability and validity (Kroenke et al., 2001; Spitzer et al., 2006). These tools enable systematic measurement of symptom severity and are often used for both screening and monitoring treatment outcomes. But patients do not have the chance to express their unique experience in using rating scales and close-ended self-report questionnaires (Grindheim et al., 2024) which is not the standard way of communicating complex psychological constructs and experiences (Armstrong, & Byrom, 2025).

Observational and behavioral assessments involve the systematic recording of overt behaviors in naturalistic or structured environments. These methods are often used in child psychology, neuropsychology, or settings where verbal self-report is limited or unreliable (Gardner, 2000). Examples include coding behavioral responses during therapy or observing interpersonal interactions. While rich in contextual information, such assessments are often resource-intensive and dependent on the skills and biases of the observer (Margolin et al., 1998).

Best-estimate assessments based on longitudinal expert appropriate data (LEAD; Eijsbroek et al., 2025a; Spitzer, 1989) involves combining all relevant data such as structured clinical interviews, rating scales, observable behaviors/markers, and clinical history from records using expert panels that review these multiple data sources to achieve a more accurate assessment. However, this approach is very costly and often infeasible in everyday practice, requiring considerable time, longitudinal follow-up, and coordination among trained professionals. As described in the LEADING guideline (a reporting guidelines for longitudinal expert appropriate data studies; Eijsbroek et al., 2025a), although best-estimate assessments represent one of the most rigorous methods for maximizing diagnostic validity, their implementation is often limited outside of specialized research settings due to the substantial resources they demand. Instead, they are primarily used as reference standards for validating other assessment tools and methods, and adapted as Longitudinal Expert Data (LED) assessment, or expert-rated assessment, offering a benchmark against which more feasible clinical assessments can be compared.

Despite the availability of empirically supported tools, their adoption in clinical practice remains inconsistent (Hunsley, & Mash, 2007). According to a recent study on clinical mental health assessment practices, many clinicians continue to rely heavily on unstructured interviews and clinical intuition (Navandi et al., under review). For example, in a multinational survey involving over 500 clinicians, 83% reported using unstructured interviews, while only 40% used structured interviews. In the same survey, 83% of clinicians also reported using rating scales to assess depression, placing them on par with unstructured interviews in terms of usage frequency. However, the way rating scales are applied in practice varies considerably. On average, clinicians used rating scales for 51% of their patients, compared to 58% for unstructured clinical interviews. Moreover, these tools were frequently used in combination with other methods rather than in isolation, suggesting that rating scales often serve as one component in a broader assessment strategy rather than as a standalone diagnostic tool. Despite their wide usage, clinicians still reported relying primarily on clinical judgment rather than structured data integration methods when interpreting rating scale results. This highlights an important gap between evidence-based recommendations and real-world application, where empirically supported instruments like rating scales may not always be used to their full potential within a systematic or standardized decision-making framework.

Furthermore, 80% of clinicians stated they relied on clinical judgment rather than statistical algorithms to integrate assessment data—even though substantial research indicates that structured and algorithmic approaches consistently outperform intuition in diagnostic accuracy and treatment planning (Grove et al., 2000; Meehl, 1954; Topol, 2019; Wong et al., 2018). In the study, “statistical algorithms” referred to tools such as clinical decision-support systems and computerized scoring models, which can systematically combine multiple sources of assessment data (e.g., symptom ratings, behavioral indicators) to guide clinical decisions. Such models typically rely on predefined rules, regression weights, or machine learning outputs to derive diagnostic probabilities or risk profiles. These methods reduce bias, enhance consistency, and often detect complex patterns that human judgment may overlook.

In sum, while traditional approaches to mental health assessment have laid a valuable foundation, contemporary clinical practice often fails to fully implement the most empirically robust methods. This gap between evidence-based tools and real-world usage highlights the need for scalable, efficient, and user-friendly assessment innovations that can enhance accuracy without compromising clinical feasibility (Kjell et al., 2024).

Core Psychometric Concepts

The credibility and usefulness of any psychological assessment depend fundamentally on its psychometric soundness, particularly its validity and reliability. These concepts are essential not only for ensuring accurate interpretations of individual results but also for establishing the scientific legitimacy of the instruments themselves. In this section, the focus will be on broader psychometric principles that underpin the validity and reliability of the assessment as a whole. As such, this section will not delve into item-level analyses typically associated with rating scales, such as item difficulty/popularity, item-total correlations or factor loadings, etc.

Validity

Validity refers to the extent to which a test or assessment measures what it claims to measure. It is arguably the most critical property of any psychological tool because without validity, even a highly consistent (reliable) test may produce results that are systematically incorrect or misleading. Next, I introduce several major types of validity that are central to psychological assessments and illustrate how they are applied throughout the empirical studies in this thesis (see also Table 1 for a summary).

Causal Validity

Causal Validity refers to whether variations in the construct being measured actually lead to variations in the test scores—implying a true causal relationship between the psychological state and the measurement outcome (Borsboom et al., 2004). In Paper III, this was tested using an experimental design known as the mood induction procedure (MIP), where different environmental settings (e.g., a church, a mall, a natural park) were used to induce distinct affective states. Language-based assessments were shown to detect these systematic changes in emotion, thereby supporting the claim that variations in underlying psychological states causally influenced language-based test outcomes.

Construct Validity

Construct Validity involves the degree to which an assessment actually measures the theoretical construct it claims to assess (Cronbach & Meehl, 1955). “It involves testing a scale in terms of theoretically derived hypotheses concerning the nature of underlying variables or constructs” (Mohaya, 2017, p. 18). This form of validity is central to psychological assessment and is particularly relevant for abstract or latent variables such as depression, and anxiety. While construct validity was an overarching aim of all studies in the thesis, it was primarily demonstrated through consistent theoretical alignment and model behavior for depression, and anxiety (see Papers I–II).

Content Validity

Content Validity refers to the degree to which the items or prompts in an assessment comprehensively cover the behaviors or constructs of interest, typically established through expert review (Fitzner, 2007). Papers I and II examined content validity by using diverse, theory-informed language prompts that targeted specific constructs like general mental health, depression, anxiety and suicidality risk. These prompts were designed to capture a broad range of psychological expressions and ensured that the language-based models were grounded in representative verbal data.

Criterion Validity

Criterion Validity refers to the extent to which a measure correlates with an external standard or outcome (Cronbach & Meehl, 1955). This thesis tested both concurrent, external, and predictive criterion validity.

Concurrent Criterion Validity: Concurrent Criterion Validity refers to the extent to which a test correlates with an established measure or criterion assessed at the same time, often when a new method is proposed as a substitute for an existing one (Cronbach & Meehl, 1955). In Papers I and II, concurrent validity was examined by examining the convergence of language-based assessments with validated rating scales (e.g., Patient Health Questionnaire-9, for depression) or expert-rated assessments of suicide risk. These assessments targeted the same time, supporting their use as valid complements or substitutes to existing methods (Cronbach & Meehl, 1955).

Concurrent Criterion Validity: Concurrent Criterion Validity refers to the extent to which a test correlates with an established measure or criterion assessed at the same time, often when a new method is proposed as a substitute for an existing one (Cronbach & Meehl, 1955). In Papers I and II, concurrent validity was examined by examining the convergence of language-based assessments with validated rating scales (e.g., Patient Health Questionnaire-9, for depression) or expert-rated assessments of suicide risk. These assessments targeted the same time, supporting

their use as valid complements or substitutes to existing methods (Cronbach & Meehl, 1955).

Predictive Criterion Validity: Predictive Criterion Validity refers to the extent to which a test can accurately forecast future outcomes that are theoretically related to the construct being measured. It is typically evaluated by administering the test and then examining how well its results correlate with relevant criteria collected at a later time (Cronbach & Meehl, 1955).

Discriminant Validity

Discriminant Validity involves measures targeting the same construct are expected to show stronger correlations with each other than with measures of different constructs (Huck, 2007, 2012). In Papers I and II, this was evaluated by examining where language-based assessment of one construct (e.g., suicidality) did not simply reflect general distress or depression. For instance, the suicidality models correlated more strongly with expert-rated suicide risk than with measures of self-harm or depressive symptoms, supporting construct specificity.

Ecological Validity

Ecological Validity refers to how well assessment findings map onto real-world behaviors and settings (Brunswik, 1949, & 2023). This thesis explored ecological validity in a preliminary way by collecting language data in semi-naturalistic settings and using free-form responses. While such methods may elicit authentic psychological expression, more comprehensive evaluations are needed to confirm ecological robustness, particularly in applied clinical settings.

Face Validity

Face Validity concerns whether a tool appears, on its surface, to measure what it is intended to measure, which is typically based on subjective judgment (Fitzner, 2007). Papers I and II, for example, examined face validity by visualizing the most predictive words and phrases for each construct—words like “hopeless” or “worthless” aligning closely with known symptoms of depression and suicidality—aiming to make the results intuitively understandable and potentially clinically meaningful.

Incremental Validity

Incremental Validity is concerned with whether a new test adds value beyond existing tools (Sechrest, 1963). Paper I demonstrated this by showing that combining multiple open-ended language formats yielded better assessments than any single format alone. Furthermore, Paper III examined whether language-based assessments could classify objective conditions following a mood induction procedure more accurately than traditional rating scales like the Positive and Negative Affect Schedule.

Taken together, these forms of validity provide a comprehensive framework for evaluating the quality of psychological assessments. Across the four papers in this thesis, these validity types are operationalized through experimental designs, clinical comparisons, and model testing, providing robust evidence for the utility of language-based approaches in psychological science.

Tabel 1.
Key types of validity

Validity	Description	Covered in paper(s)
Causal Validity	Causal Validity refers to when “a test is valid for measuring an attribute if (a) the attribute exists and (b) variations in the attribute causally produce variation in the measurement outcomes.” (Borsboom et al., 2004, p. 1061).	III
Construct Validity	Construct Validity “involves testing a scale in terms of theoretically derived hypotheses concerning the nature of underlying variables or constructs [Pallant, 2010]. (Mohaya, 2017, p. 18).	I & II
Content Validity	Content Validity concerns “an exhaustive review by an expert panel to decide whether the types of questions (items) adequately cover the behavior that you are interested in measuring.” (Fitzner, 2007, p. 776).	I & II
Criterion Validity	***Concurrent Criterion Validity.*** Concurrent validity refers to when “the test score and criterion score are determined at essentially the same time…. Concurrent validity is studied when one test is proposed as a substitute for another (for example, when a multiple-choice form of spelling test is substituted for taking dictation), or a test is shown to correlate with some contemporary criterion (e.g., psychiatric diagnosis).”; Cronbach & Meehl, 1955, p. 2).	I & II
	External Criterion Validity. External Validity concerns what populations, settings, treatment variables, and measurement variables a test can be generalized to (Campbell. 2015, p. 5).	I & II
	Predictive Criterion Validity. Predictive Validity refers to when one “administers the test, obtains an independent criterion measure on the same subjects, and computes a correlation. If the criterion is obtained some time after the test is given, [it is referred to] predictive validity”. (Cronbach & Meehl, 1955, pp. 1-2)	-
Discriminant Validity	Discriminant Validity means “the correlations between … measures should be larger than their respective correlations to the other questionnaires because of the latters’ intention to measure other constructs.” (Huck, 2007, 2012, p. 84).	I & II
Ecological Validity	Ecological Validity means “the response variable is validated not against another response but against an antecedent stimulus variable (distal or proximal) established within the same sample of situations” (Brunswik, 1947, p. 174).	-
Face Validity	Face Validity concerns “a subjective judgment of whether the tool or question is a good measure or not. ”(Fitzner, 2007, p. 776).	I, II & III
Incremental Validity	Incremental Validity is concerned with the concept that “the test will add to or increase the validity of predictions made on the basis of data which are usually available.”(Sechrest, 1963)	I & II

Reliability in Psychological Assessment

Reliability refers to the consistency and stability of a psychological assessment tool (under stable conditions), across raters or time (e.g., 2 weeks of test-retest), or within the test itself (Matheson, 2019). It is a fundamental psychometric property because an instrument that produces erratic or unstable results cannot be trusted to yield accurate information about an individual's psychological state. In both clinical and research settings, high reliability is essential for drawing meaningful inferences from test scores, monitoring change over time, and ensuring that assessment results are not simply artifacts of measurement error.

There are several types of reliability, each capturing a different aspect of measurement stability. They provide important context for interpreting the robustness of language-based assessments developed across the included studies. Together, these reliability concepts offer a framework for evaluating measurement consistency. While language-based assessments structurally differ from traditional tools, this thesis offers preliminary evidence that they may meet some reliability standards in research settings—particularly through prospective evaluation and test–retest designs—ensuring that the models are valid, consistent and dependable over time and across samples (see Table 2 also for an overview of reliability concepts). Further studies, particularly in clinical populations and real-world contexts, are needed to confirm their robustness in the future.

Internal Consistency Reliability

Internal Consistency Reliability refers to how well the items within a test or measure assess the same underlying construct (Dunn et al., 2014). While this form of reliability typically applies to multi-item rating scales, it is relevant here conceptually. In language-based assessments, internal consistency would translate into the coherence of linguistic features that consistently signal a given psychological construct. Although not formally assessed in this thesis, future work could explore internal linguistic homogeneity across different language response formats to evaluate how well semantic patterns converge on specific constructs like depression or anxiety (Henson, 2001).

Interrater Reliability

Interrater Reliability refers to the extent to which different raters or observers produce consistent scores when evaluating the same target (Zhao et al., 2022). This form of reliability was explicitly examined in Paper II, where expert-rated suicide risk assessments were made by expert clinicians.

Parallel-Forms Reliability

Parallel-Forms Reliability assesses whether two different forms of a test, which are intended to be equivalent, yield similar results (Malau-Aduli et al., 2012). While not directly tested in this thesis, this concept is indirectly relevant. The studies employed different response formats—such as writing words, phrases, or full narratives—that could be conceptualized as parallel expressions of similar psychological constructs. Evaluating equivalence across these formats remains an important avenue for future research, particularly for determining how format selection influences interpretability and reliability (Lord, 1983).

Prospective Reliability

Prospective Reliability captures the performance of a model or test when applied to a new set of participants, often in a pre-registered design (Kjell et al., 2024). This form of reliability was central to Paper I, which followed a sequential evaluation with model pre-registration (SEMP, see Chapter 4 for a detailed explanation) framework. Models were first trained on development samples, pre-registered, and then tested on independent prospective samples. The models showed consistent performance across held-out datasets under research conditions (Kjell et al., 2024). However, their predictive utility in clinical or high-stakes decision-making contexts remains to be validated.

Split-Half Reliability

Split-Half Reliability evaluates internal consistency by splitting a test into two segments and analyzing the correlation between the scores of each segment (Cooper, 2023). This method was not used in the present studies, as language-based assessments do not rely on traditional itemized scales. However, future work could investigate split-half methods by randomly dividing participants' linguistic features or responses and evaluating model agreement across splits, potentially providing an adaptation of this classic reliability concept.

Test–Retest Reliability

Test–Retest Reliability evaluates the temporal stability of an assessment by comparing scores obtained at different time points where no change should really have occurred (Trimble, 1943). In Paper I, this form of reliability was tested over a two-week interval. Results showed moderate to high consistency in model outputs. These findings demonstrate that language-based assessments can produce stable and repeatable results over time, an essential quality for tools used in monitoring and longitudinal assessment.

Tabel 2.
Key types of reliability

Reliability	Description	Covered in paper(s)
Internal Consistency Reliability	Internal Consistency Reliability indicates the degree to which test items assess the same construct, ensuring that a composite score can be meaningfully interpreted as a unified reflection of the intended attribute. (Henson, 2001).	-
Interrater Reliability	Interrater Reliability refers to the concept that "measurement of the extent to which data collectors (raters) assign the same score to the same variable is called interrater reliability" (McHugh, 2012, p. 276)	II
Parallel-Forms Reliability	Parallel-Forms Reliability assesses whether two different forms of a test, which are intended to be equivalent, yield similar results (Malau-Aduli et al., 2012).	-
Prospective Reliability	Prospective Reliability involves "apply the pre-registered models to a "prospective holdout test" sample consisting of new participants. The base hypothesis for the evaluation phase is that the models trained during the development phase will continue to predict their intended outcomes on the unseen data." (Kjell et al., 2024, p. 6)	I, II (only holdout), & III (not pre-registered)
Split-Half Reliability	Split-Half Reliability means "Divide the items into two halves and correlate scores on these two halves together" (Cooper, 2023, p. 149)	-
Test-Retest Reliability	Test-Retest Reliability concerns "tests at intervals with the same test or with equivalent tests." (Trimble, 1943, p. 481)	II

Summary

Psychological assessments are essential tools in both clinical and research contexts. Their utility hinges not only on what they measure but on how accurately and consistently they do so. In this chapter, I have reviewed the core qualities that define a robust assessment: validity, which ensures that an instrument measures what it intends to measure, and reliability, which ensures that it does so consistently across time, raters, and samples.

These foundational psychometric principles, developed over decades of research, continue to guide the evaluation of newer forms of assessment, including those based on natural language. As language-based assessment methods become more common in psychological science, it is critical that they be held to the same standards of scientific rigor, while also addressing new challenges unique to open-text data—such as feasibility, interpretability, and data privacy.

The next chapter will lay the theoretical groundwork for understanding language not just as a medium of communication, but as a measurement tool in its own right. We will explore the conceptual underpinnings of language-based approaches, drawing from philosophy, cognitive science, and psychometrics to frame language as a vehicle for psychological insight.

Chapter 3: Theoretical Foundations of Language-Based Psychological Assessments

Introduction: Why Language Matters in Psychological Science

Language is a central feature of the human condition. It is the primary means through which people express their thoughts, emotions, intentions, and experiences. From everyday conversation to therapy sessions and diagnostic interviews, language provides a direct and nuanced channel for revealing psychological life. As such, it is not merely a medium of communication but a data-rich behavioral expression that carries significant psychological meaning (Boyd, & Pennebaker, 2015; Boyd, & Schwartz, 2021; Jackson et al., 2022).

In clinical contexts, language is the foundation of most interactions. Clinicians rely on clients' verbal reports to assess symptoms, explore life histories, evaluate progress, and establish therapeutic rapport. Diagnostic criteria themselves are often based on linguistic expressions of distress (e.g., "I feel worthless" or "I can't stop worrying") (Berry-Blunt et al., 2021; Bredström, 2019). Whether structured through interviews or elicited through open-ended prompts, the language used in clinical conversations is a rich source of evidence about internal states. This makes verbal behavior a potentially informative target in psychological assessment (Boyd, & Schwartz, 2021).This chapter outlines the theoretical foundations underlying the use of language as a tool for psychological assessment. We begin by examining how language reflects mental processes and the advantages it offers as a measurement tool. We then review behavioral and contextual models of language use and conclude with a discussion of how modern AI techniques are poised to transform assessment science.

Language and Context

The importance of context in language use is well-established across psychological and computational traditions. For example, transformer-based models such as BERT and GPT dynamically adjust the word embedding—the numeric representation of a word—based on the linguistic context in which the word appears (Devlin et al., 2019; Vaswani et al., 2017). This represents a significant advancement over earlier models that assigned fixed embeddings to each word, regardless of usage. In these newer architectures, the same word (e.g., “cold”) will have a different representation depending on whether it appears in a medical context (“I have a cold”) or a psychological or interpersonal one (“She’s been so cold to me lately”). While this ability to model local context is a crucial step forward, it can be taken further.

Models must also account for broader, extralinguistic contexts—such as the identity of the speaker, the social relationship between speaker and listener, the medium of communication, and the physical or emotional setting—factors that meaningfully shape how language is produced and interpreted. Boyd et al. (2021) argue that words are never isolated signals but are embedded within nested layers of discourse, shaped by time, place, audience, and interactional purpose. Meaning is not merely contained in individual words, but constructed through the broader flow of conversation and the social dynamics in which it unfolds. Building on this, Kjell et al. (2024) emphasize that language is shaped by who is speaking, to whom, where, and how—with differences emerging across demographic groups, social roles, physical settings, and response modalities. For instance, whether someone is responding to a chatbot, a clinician, or a virtual avatar may profoundly affect their willingness to disclose sensitive information. Yet despite this rich understanding of language as situated and socially embedded, the models in the thesis treat language primarily as decontextualized input—stripped of interpersonal, situational, and emotional context. This limits their ability to capture how meaning is shaped by narrative, silence, metaphor, or strategic self-presentation, especially in clinical settings where expression is often layered and affectively charged.

This might be problematic in clinical contexts, where language is often not a straightforward report of internal states but a complex, strategic, and emotionally laden form of expression. Research in psychotherapy and clinical linguistics has long emphasized that how something is said—or not said—can carry more diagnostic or therapeutic significance than its literal content (Frank & Frank, 2025; Levitt et al., 2018). Patients may use metaphor to externalize pain (“a weight on my chest”), or deflection to manage emotional vulnerability, and silences can signal affective regulation, resistance, or trust dynamics (Levitt, 2001). Such features are central to meaning-making in therapeutic settings, yet they are difficult to quantify and are not easily captured by models optimized for predictive accuracy alone. By treating language primarily as a feature vector for classification, the current models may miss precisely those aspects of verbal expression that clinicians often find most meaningful—those that reveal not just what a person feels, but how they are navigating, containing, or communicating their distress.

In this regard, the analyses presented in the current thesis represent a limitation. While the models developed here take steps toward using open-ended language in assessment, they rely solely on text and do not incorporate multimodal or temporally embedded information. More recent models are beginning to address this. For example, WhiSPA (Rao et al., 2025) integrates audio features such as prosody and voice quality with language content to detect mental health states more effectively. Similarly, video-based systems now analyze facial expressions and gestures to enrich understanding of affective dynamics (Soleymani et al., 2017 for a survey). Other approaches, like the HaRT model (Soni et al., 2022), track an individual's language history over time, tailoring embeddings to the speaker's unique linguistic patterns and psychological baseline. These approaches move toward more ecologically valid and personalized models by taking fuller account of how, when, and by whom language is produced. Future work will need to engage with these advancements in order to better capture the layered, multimodal, and situated nature of psychological expression. Although these perspectives highlight language as socially and contextually embedded, in this thesis they mainly served as methodological justifications, a limitation I return to in Chapter 7.

Conceptualizing Language in Psychological Terms

Language is not only a tool for communication but a reflection of what individuals attend to, think about, and feel (Xintong, & Xiaofei, 2024). Words reveal the focus of attention, the structure of thought, and the emotional tone of experience. When someone speaks or writes, they externalize internal states, making language a valuable proxy for psychological processes.

This view has deep theoretical roots. Early thinkers like Francis Galton (Galton, 1884, in Uher, 2013) proposed that the words people use encode social and psychological needs, while Ferdinand de Saussure (Saussure, 1962, in Brandt, 2022) emphasized that language reflects the underlying structure of the human mind. These foundational ideas continue to influence how language is interpreted in modern psychological science.

Today, language is increasingly recognized as both a product and a mirror of psychological activity. It arises from cognitive and emotional processes, and in turn, shapes how experiences are structured and remembered (Lindquist, & Gendron, 2013). As such, verbal behavior provides a window into the mind—one that can now be systematically analyzed through computational methods to reveal patterns linked to mental health, personality, and behavior (Eichstaedt et al., 2018).

Language as Situated Psychological Expression. While computational models allow language to be treated as structured data, psychological theory emphasizes that language is more than a stream of words—it is a situated act of expression. Narrative theory highlights that people make sense of their lives and communicate emotions through stories that integrate past, present, and imagined futures

(McAdams, 2001). Discourse psychology further situates language within interpersonal and cultural contexts, showing how meanings are constructed in dialogue and shaped by social positioning (Bowers, 1988). Clinical process models underscore how verbal expression is entwined with therapeutic mechanisms such as emotional disclosure, alliance formation, and defense (Greenberg & Pascual-Leone, 2006). Finally, affective science demonstrates how emotional states are encoded linguistically, for instance through valence- and arousal-related expressions, which bridge subjective experience and observable behavior (Barrett, 2017). Together, these perspectives clarify that language-based assessments should not be seen as abstract pattern recognition but as reflections of meaning-making processes that are psychological, relational, and embodied. By aligning computational approaches with these traditions, this thesis situates language-based assessments within a broader theoretical understanding of how individuals express and regulate internal states.

Advantages and Disadvantages of Language as a Measurement Tool

Advantages

Language offers several potential advantages as a research tool in psychological assessment. Unlike rating scales, which constrain responses to predefined categories, natural language enables rich, flexible, and multidimensional expression (Kjell et al., 2024).

Range. Language allows individuals to describe both subtle feelings and extreme psychological states. From mild unease to profound despair, the breadth of vocabulary provides the means to express a wide spectrum of mental health experiences.

Resolution. It supports fine-grained distinctions between similar states. For example, someone can distinguish between being “worried,” “anxious,” or “panicked,” offering more precise insights than a single Likert item could convey.

Dimensionality. Natural language permits the expression of multiple psychological dimensions at once. A person might describe feeling “nervous but hopeful,” capturing complex emotional blends that are difficult to encode in fixed-scale items.

Openness. Language is inherently open-ended and adaptable. People can respond in ways that reflect their cultural background, personal context, and unique perspectives, allowing for a more personalized and ecologically valid assessment.

Information Content. Quantitatively, open responses tend to contain more information—measured as entropy or diversity—than rating scales (Kjell et al., 2022; Kjell et al., 2024; Paper I). This higher information density may offer richer distinctions under research conditions.

Language as a Behavioral and Situational Marker

While language reflects internal psychological processes, it is also a form of observable behavior—situated, context-dependent, and shaped by motivational and social forces. From this behavioral perspective, language is not only an output of cognition but also a dynamic act through which individuals pursue goals, navigate social environments, and enact habitual patterns (Boyd, & Schwartz, 2021). People use language to signal affiliation, assert identity, request support, and manage impressions (Roberson et al., 2024). In this way, language becomes a tool of action and interaction—an observable trace of internal motives and external demands.

The meaning of language is also inherently contextual (Demszky et al., 2023). A word or phrase does not carry a fixed psychological interpretation outside its grammatical, semantic, and situational surroundings (e.g., Fig. 1 in Boyd, & Schwartz, 2021). For example, saying "I'm fine" may signal contentment in one situation and emotional withdrawal in another, depending on tone, relational context, and recent events. Thus, to interpret language accurately in psychological assessments, we must account for syntactic structure, lexical ambiguity, and discourse-level meaning. Contextual nuances—such as irony, negation, or emotional intensity—can critically change how a response should be understood.

Verbal expression is further shaped by the social and situational environment in which it occurs. Language produced in therapy differs from that in casual conversation or anonymous surveys. Individuals tailor their word choices based on audience, perceived norms, and emotional safety (Flusberg et al., 2024). This responsiveness makes language a sensitive indicator of psychological state, but it also introduces variability that must be interpreted carefully (Mangalik et al., 2024). Factors such as power dynamics, cultural background, and the immediacy of experience all influence how thoughts are verbalized.

To synthesize these elements, the interactionist model provides a useful framework (Judge, & Zapata, 2015; Tett, & Burnett, 2003). It posits that behavior—including language—emerges from the interaction between the individual (traits, emotions, intentions), the social context (roles, norms, relationships), and the situational constraints (task, environment, stressors). This model supports a broader interpretation of language: not just as a reflection of mental content, but as a behavior shaped by context—yet how this translates into reliable clinical inference remains to be tested. When applied to assessment, this means that understanding someone's language requires understanding not just what they say, but why, how, and in what situation they say it.

Limitations

Susceptibility to Social Desirability Bias. Despite the openness and richness of language-based responses, participants may still tailor their language to conform to perceived social norms or expectations. Similar to rating scales, natural language is vulnerable to social desirability bias, where individuals present themselves in a more favorable light (McDonald, 2008). This can distort the authenticity of the language data, especially when discussing stigmatized topics like mental health, thus potentially undermining the validity of assessments based on open-ended responses.

Limitations Due to Lack of Insight. Another challenge inherent to language-based assessments is that their quality fundamentally depends on individuals' self-awareness and capacity for introspection. People vary in how accurately they can recognize, understand, and articulate their psychological states (Hull et al., 1988). Prior research has shown that individuals often have limited insight into their own emotions and behaviors (Kenny et al., 1994), which can constrain the depth and accuracy of the language they produce, regardless of how sophisticated the AI models analyzing the language are.

Vulnerabilities Related to Introspection and Expressive Ability. Language-based assessments also presuppose a minimum level of expressive ability. Some individuals—due to cognitive impairments, developmental differences, or educational background—may struggle to articulate complex or nuanced psychological experiences (Odermatt et al., 2025). This variation can introduce systematic biases, where assessments may perform better for verbally skilled individuals while underestimating psychological states in those who have greater difficulty expressing themselves. Thus, while offering expressive flexibility, language-based assessments may introduce disparities based on introspective and linguistic abilities, especially in real-world or clinical settings.

Elicited Versus Naturally Occurring Language in Psychological Assessment

A central distinction in language-based psychological assessment lies in the origin of the language data: whether it is probed—elicited through structured prompts (Kjell et al., 2019) —or already existing—naturally occurring in everyday contexts such as social media (e.g., Eichstaedt et al., 2015), therapy transcripts (e.g., Tanana, 2016), or diary entries (e.g., Linton et al., 2021). Probed language refers to verbal responses generated in direct response to specific questions or instructions (e.g., "Describe how you have been feeling lately"), whereas already existing language captures spontaneously produced expressions not intended for assessment purposes.

This distinction matters because it influences the interpretability, standardization, and ecological validity of the resulting assessments.

Probed language offers high levels of experimental control, comparability across participants, and alignment with theoretical constructs (Demszky et al., 2023; Jackson et al., 2022). In contrast, already existing language is ecologically valid, reflecting the authentic contexts in which people naturally express themselves. However, it is less standardized, which can complicate comparisons across individuals or time points.

In sum, both probed and naturally occurring language offer different strengths and limitations for language-based research. Rather than viewing them as mutually exclusive, they can be seen as complementary sources of information (Jackson et al., 2022). Future work may benefit from hybrid models that incorporate both types of language to maximize the trade-off between control and ecological validity, enabling richer and more flexible approaches to understanding mental health through language.

From Theory to Technology: Foundations of AI-Based Language Analysis

The emergence of natural language processing has reshaped how psychological language can be systematically analyzed. In earlier decades, language was mostly treated qualitatively or through basic word counts (e.g., linguistic inquiry and word count, LIWC, Pennebaker et al., 2001), limiting the resolution and interpretability of findings. However, the growing integration of computational methods in psychology has enabled researchers to analyze free-text responses at a scale and depth that was previously unattainable (Feuerriegel et al., 2025). This evolving paradigm—often referred to as computational language-based assessment (Kjell et al., 2022)—opens new avenues for measuring psychological constructs using the richness of individuals' own words.

One early theoretical model underpinning computational language analysis is the "words-as-attention" framework. This model suggests that the words people choose reflect what they are paying attention to, cognitively and emotionally (Boyd, & Schwartz, 2021). For example, a person who uses many negative emotion words may be attending to distressing experiences or interpreting events through a negative lens. This view has guided early and recent works in language and mental health, including research linking word frequencies to depression, anxiety, and well-being (Kaźmierczak et al., 2024; Pennebaker, 1997). However, while this approach provides useful psychological insights, it has clear limitations (Boyd & Schwartz, 2021). It treats words largely in isolation, disregarding syntax, context, and the interplay between words in forming meaning. Words like “fine,” “nothing,” or “overwhelmed” can have vastly different implications depending on their linguistic and situational context—something traditional word-count methods cannot easily capture. For instance, the word “nothing” could signal contentment (“Nothing’s wrong, everything’s good”) or deep distress (“Nothing matters anymore”), depending on how and when it is used.

Advances in natural language processing, particularly the development of transformer-based large language models (LLMs) such as BERT (e.g., Bidirectional Encoder Representations from Transformers, in Devlin et al., 2019), RoBERTa (e.g., Robustly optimized BERT pretraining approach, in Liu et al., 2019), and GPT (e.g., Generative Pre-Training, in Radford et al., 2018), have addressed many of these limitations. These models learn language representations by considering the full context in which a word appears, rather than just its surface frequency. Through self-attention mechanisms, they dynamically weigh the importance of each word relative to its surrounding words, enabling more accurate understanding of meaning. This capacity for contextual modeling has significantly improved the ability of machines to interpret emotion (e.g., Tanana et al., 2021), cognition (e.g., Radford et al., 2023), and interpersonal nuance (e.g., Kjell et al., 2021) in text.

At the heart of these models is the concept of embeddings—high-dimensional numerical representations of words, phrases, or entire texts (Demszky et al., 2023). Embeddings encode semantic similarity such that texts with similar psychological meanings are placed close together in the embedding space. Unlike earlier representations (e.g., one-hot vectors which use binary variables for the presence of each word, Cerda et al., 2018), modern embeddings are both dense and context-sensitive, allowing them to preserve subtle affective or cognitive distinctions between responses. When combined with machine learning, these embeddings can be used to assess psychological outcomes or construct scores for constructs relating to mental health.

Together, these technological advances offer a promising foundation for exploratory language-based assessments. They aim to move beyond keyword-based systems by modeling language contextually.

Summary

This chapter has laid the theoretical groundwork for understanding language as both a reflection of internal psychological states and an observable behavioral expression shaped by social and situational context (Boyd, & Schwartz, 2021). The chapter began by examining the central role of language in psychological science—how it captures attention, emotion, cognition, and motivation. Then the chapter explored the unique advantages of language as a measurement tool, from its expressiveness and dimensionality to its openness and information richness. Importantly, the chapter highlighted the difference between elicited (probed) and naturally occurring language and how each serves different scientific purposes. Finally, this chapter traced the technological evolution from early word-count methods to modern large language models capable of context-aware and fine-grained psychological analysis.

Theoretically, this chapter moves beyond the view of language as merely an index of attention or a static set of words. Instead, it embraces a more expansive behavioral framework: one that recognizes language as a motivated, goal-directed act embedded in specific social, cultural, and interpersonal contexts (Boyd, & Schwartz, 2021). This perspective aligns with interactionist and ecological models in psychology and opens the door to a richer understanding of how and why people express themselves as they do.

Looking ahead, language-based assessments may hold promise for psychometric research (Bhatia et al., 2022; Boyd, & Schwartz, 2021; Demszky et al., 2023; Eichstaedt et al., 2015; Kjell et al., 2024). They are rich, flexible, and capable of capturing nuances that traditional instruments may miss. However, to fulfil this promise, they must meet the same scientific standards of validity, reliability, and interpretability that guide conventional tools. This includes not only assessment accuracy but also transparency, fairness, and clinical applicability.

The following chapter marks a shift from theoretical foundations to methodological application. It details the implementation, evaluation, and validation of language-based assessments within the scope of this thesis, incorporating both established psychometric principles and contemporary approaches, including model pre-registration, experimental validation, and open-science infrastructure. This methodological foundation is critical for evaluating whether language-based models can ultimately become trustworthy, robust, and potentially applicable beyond research contexts.

Chapter 4: Methods Used in the Studies

An overview of methodological approaches

The four papers included in this thesis employ a diverse set of methodological approaches that reflect the interdisciplinary nature of language-based psychological assessments (Mangalik et al., 2024). At the core of these studies is the use of artificial intelligence—particularly large language models—to transform open-ended natural language responses into computational representations suitable for psychological evaluation. Across the papers, language data is processed and visualised using natural language processing techniques modeled using machine learning algorithms with robust cross-validation procedures, and evaluated through established psychometric frameworks (see Ch. 2 for details). The methodological strategies also encompass advanced validation procedures, including the sequential evaluation with model pre-registration framework, which supports transparent and rigorous model testing (Kjell et al., in progress). In addition, the thesis incorporates both correlational, longitudinal and experimental designs, including the use of expert-rated clinical ratings via the longitudinal expert data (LED) assessment procedure (Eijsbroek et al., 2025a) and mood induction procedures to evaluate causal validity (Borsboom et al., 2004). Finally, the methodological work is grounded in principles of open science, exemplified by the development and dissemination of reusable models through the language-based assessment model (L-BAM) Library and supporting software tools (Kjell et al., 2023). The following sections offer a more detailed overview of each methodological component described above, moving from general approaches to specific techniques applied in the included studies.

Language-Based Assessments Using AI

The core analytical approach across all four papers involves the use of language-based assessments, which leverage artificial intelligence to quantify individuals' psychological states based on their natural language responses. Specifically, Language-based assessments utilise large language models to transform written language into high-dimensional numerical representations known as word embeddings (Demszky et al., 2023). These embeddings capture the semantic content of words and phrases based on their contextual use, allowing for subtle emotional, cognitive, and behavioral markers to be identified and analyzed computationally (Kjell et al., 2024).

Once text is converted into embeddings, models are trained using machine learning algorithms to assess various psychological constructs—including depression, anxiety, suicidality risk, well-being, and condition assignment in experimental settings. These assessment models are evaluated using rigorous cross-validation procedures, most commonly k-fold cross-validation (Kjell et al., 2023), to ensure that results generalize beyond the training data. This process provides robust estimates of out-of-sample performance and guards against overfitting.

In addition to assessment modelling, natural language processing methods are used to visualize and interpret the linguistic content of participants' responses (Eijbroek et al., 2025b; Sikström et al;, 2025). Techniques help illustrate which words or expressions are most indicative of particular psychological states. These visualizations not only aid in interpreting model decisions but also serve as visualization aids that may enhance interpretability and transparency of the assessments.

Together, these AI-based methods provide a research-driven alternative to closed-ended survey instruments by using the richness of everyday language. The use of embeddings, assessment modeling, and visual analytics forms the technical foundation of the language-based assessment framework employed throughout this thesis.

Best-estimate Assessment through Longitudinal Expert Assessment of Appropriate Data (LEAD)

The expert ratings used for validating models in Paper II followed the longitudinal expert appropriate data (LEAD) approach (Eijbroek et al., 2025a), as outlined earlier in Chapter 2, but was adapted to longitudinal expert data (LED) assessment. Here, the process is applied to generate clinically grounded outcomes, offering an exploratory benchmark for assessing model outputs, while acknowledging that expert ratings differ from diagnostic consensus methods.

Sequential Evaluation with Model Pre-registration

To ensure methodological rigor, the studies in Paper I and II employed the sequential evaluation with model pre-registration framework (Kjell et al., in progress), involving developing models on a training set, pre-registering all modeling choices, and then evaluating performance on an unseen holdout or prospective sample. The sequential evaluation with model pre-registration approach aims to minimize overfitting and support initial confirmation of model performance across different psychological constructs.

Mood Induction Procedure: Experimental (Causal validity) Design

To evaluate the causal validity of language-based assessments —that is, their sensitivity to detecting systematic changes in affect resulting from external manipulations—this thesis incorporates an experimental paradigm based on a mood induction procedure (Marcusson-Clavertz et al., 2019). The mood induction procedure is a well-established method in psychological research for eliciting changes in participants' emotional states in a controlled and replicable way (Westermann et al., 1996).

In the study included in Paper III, participants were randomly assigned to experience one of three distinct environmental settings designed to induce different affective states: a church, a shopping mall, or a park. These conditions varied in their sensory, social, and symbolic features and were selected to elicit unique affective profiles. The assigned condition served as an objective criterion—a known "ground truth" against which the sensitivity of different assessment methods could be tested. The performance of language-based assessments in classifying participants' assigned condition based on their language responses was directly compared to that of traditional rating scales . This design allowed for a test of measurement responsiveness, providing evidence for whether an assessment tool may detect induced changes in psychological states when those changes are experimentally induced.

Open Science and the L-BAM

Open science refers to the movement toward greater transparency, accessibility, and reproducibility in scientific research (Hales, et al., 2019). By sharing data, code, and tools, open science practices aim to accelerate discovery, improve collaboration, and promote cumulative progress across disciplines (Dudda et al, 2025). In psychological assessment, open science is especially vital to ensure that newly developed tools—particularly those powered by AI—can be independently verified, compared, and adapted for use across different populations and settings (e.g., Yawer et al., 2023).

This thesis integrates open science principles throughout its empirical work, including the development of the L-BAM Library and the *textAssess()* tool—both of which facilitate the practical dissemination and application of validated language-based models. The L-BAM Library offers a standardized infrastructure for sharing pre-trained assessment models, accompanied by documentation that supports reproducibility, transparency, and ease of use for researchers in psychology and related fields.

In sum, this thesis embodies open science values as an integral methodological foundation—making validated tools freely available, enabling independent research replication and exploration of AI-powered assessments. Specifically, L-BAM encourages researchers to conduct independent validation of models and to test their applicability in new populations, languages, and research contexts. By doing so, it supports cumulative science and safeguards against overfitting and unwarranted generalizations (Open Science Collaboration, 2015), supporting responsible and exploratory use of AI models in psychological research.

Limitations

While the methods introduced in this thesis—including language-based assessment workflows, the sequential evaluation with model pre-registration validation framework, longitudinal expert data (LED) assessment-based evaluations, mood induction experiments, and open dissemination via the L-BAM Library—collectively strengthen the transparency and methodological rigor of language-based assessments for research purposes, they are not without limitations. Each methodological innovation brings specific challenges that warrant consideration. For instance, language-based assessments depend on the quality and variability of participants' language, which can be shaped by introspective ability, literacy, and social context. The generalizability of sequential evaluation with model pre-registration findings relies on the representativeness of the development samples and the appropriateness of the pre-registered choices. Longitudinal expert data (LED) assessment ratings, while clinically robust, are inherently influenced by expert interpretation and may vary across raters. mood induction procedure, although ecologically motivated, may not evoke consistent affective responses across individuals or translate easily to real-world settings. Finally, open sharing of models—while critical to scientific openness—introduces ethical and practical concerns related to data privacy, appropriate use, and long-term maintenance of quality. Addressing these challenges will require ongoing methodological development, careful contextual application, and a commitment to continuous validation across diverse settings and populations.

Chapter 5. Summary of the Research Papers

Paper I

Natural Language Response Formats for Assessing Depression and Worry with Large Language Models

Aims

The overarching aim of Paper I was to investigate whether varying the degree of openness in language-based response formats influences the validity and reliability of psychological assessments, particularly in the context of assessing depression and anxiety—two of the most prevalent and diagnostically relevant mental health conditions globally.

Recognising that traditional mental health assessments rely heavily on closed-ended rating scales, which often fail to capture the richness and nuance of individuals' subjective experiences, the study focused on developing and testing a set of systematically varied response formats. These formats ranged from relatively constrained to fully open, designed to enable participants to express their mental health states using natural language. Specifically, four response formats were developed and evaluated:

1. **Select words format** – participants selected words from a predefined list. This format combined a structured question with a language-based answer, making it potentially useful in settings that require both standardization and speed.
2. **Write words format** – participants were prompted to generate their own descriptive words that best captured their experiences with depression or worry. This format preserved brevity with language-based assessment models demonstrated more encouraging individuality and expressiveness.
3. **Write phrases format** – participants were asked to write short phrases describing their psychological states, which added syntactic structure to the language, allowing for more contextual nuance than single words.

4. **Write texts format** – the most open-ended format, where participants freely wrote complete sentences or short narratives about their mental health experiences. This format was expected to offer the richest linguistic information due to its complexity and breadth.

The work also addressed an important methodological gap by systematically comparing different natural language response formats and quantifying their respective strengths and limitations using rigorous psychometric criteria and advanced AI techniques.

Methods

Participants. Participants were recruited from Prolific (Palan & Schitter, 2018), a well-validated online platform specialized in recruiting participants to do psychology experiments and surveys. The final development dataset included responses from 963 individuals. For the evaluation phase, a pre-registered prospective test sample of 145 new participants was collected to assess generalizability. Participants were fluent in English and aged 18 or older.

Procedure and Measures. Participants were asked to respond to the standardized language-based assessment prompts assessing either depression or worry/anxiety in random order. After submitting their open-ended responses, participants also completed traditional rating scales, including two validated rating scales for depression and another two for anxiety/worry. These scale scores served as the criterion for evaluating the accuracy of the language-based models. Lastly they answered a brief survey on demographics and clinically relevant information such as number of sick-leave days over the last month/year. In the prospective test sample, participants completed the same procedure using the response format assigned in the development phase, allowing for direct replication and validation of the pre-registered models.

Results

The study provided a comprehensive evaluation of four distinct natural language response formats—select words, write words, write phrases, and write texts—for assessing depression and anxiety. The evaluation spanned both development and prospective validation phases, yielding insights into multiple forms of validity and reliability.

Concurrent (Criterion) Validity. Across all four response formats, the language-based assessment models demonstrated moderate to strong concurrent validity, as evidenced by high correlations with corresponding traditional rating scale scores. Specifically, correlations between model assessments and scale scores ranged from *r* = .59 to .77, suggesting that each format was capable of reflecting individuals' self-reported levels of depression and anxiety to a substantial degree. These correlations approached the theoretical upper bounds of validity, which are typically constrained by the internal reliability of the scales themselves.

Concurrent (Criterion) Validity Across Sample Sizes. An analysis investigated whether increasing the training sample size led to improved model accuracy. Results indicated that model performance did indeed benefit from larger training datasets, with accuracy metrics climbing steadily as the amount of training data increased. This finding underscores the importance of robust, adequately powered samples in developing generalizable language-based assessment models, especially for capturing nuanced psychological phenomena.

Incremental Validity. When different response formats were combined, the resulting composite models yielded even higher levels of assessment accuracy. Specifically, aggregating information from multiple formats led to a correlation of *r* = .83 with corresponding rating scales—approaching the scales' own reliability. This result supports the hypothesis that different response formats offer partially non-overlapping insights and that combining them leverages complementary strengths. The additive benefit of including both constrained (e.g., select words) and open-ended (e.g., write texts) formats suggests that a multi-format assessment strategy may yield the most accurate and comprehensive understanding of individuals' mental health. However, combining all formats is at the cost of time.

Face Validity. To assess face validity, the study generated word-level visualisations showing which words were most predictive of depression versus anxiety, across the different formats. These visualizations showed that the language used by participants aligned with established symptomatology and theoretical expectations. For example, terms like "worthless," "sad," and "lonely" were reliably associated with higher levels of reported depression. These linguistically coherent and contextually relevant associations offer support for the interpretability and face validity of the models, which is critical for clinical utility.

Discriminant Validity. The study also tested whether models trained on depression versus anxiety/worry responses could meaningfully distinguish between these two constructs. Results supported this discriminant validity: although depression and anxiety are often comorbid and correlated, the models captured meaningful linguistic distinctions. Depression was associated with more words of diminished interest or pleasure, whereas anxiety and worry were more often characterized by descriptors of excessive anxiety and worry along with difficulty to control the worry. This differentiation highlights the capacity of language-based assessments to parse closely related but theoretically distinct constructs.

Prospective Reliability. When tested on a held-out prospective sample of participants (*N* = 145), the pre-registered models retained high validity. They met and exceeded the pre-specified hypothesis that models would yield at least moderate correlations (*r* > .50) with total scale scores. In fact, the correlations remained almost within the range observed during development (*r* = .59–.77), demonstrating strong generalizability and robustness across samples.

Test-Retest Reliability. To assess the temporal stability of the language-based models, test-retest correlations were calculated over a two-week interval. The results demonstrated moderate reliability of language-based assessment (LBA), while the rating scales were higher, indicating that the language-based assessments produced consistent scores for the same individuals across time. This supports the models' potential for use in longitudinal assessments, treatment monitoring, and follow-up evaluations.

External (Criterion) Validity. The models also exhibited strong external validity. Language-based assessments showed significant correlations with self-reported behavioral indicators, including sick leave and healthcare visits due to mental health problems. In 9 out of 12 comparisons, these language-based assessments outperformed traditional rating scales in converging with such external outcomes.

Information Content. As expected, the amount of psychological information contained in responses increased with the openness of the format. The write phrases formats provided the highest information richness, allowing for nuanced and idiosyncratic expression of symptoms. Conversely, the select words format offered the least information, though still useful in constrained settings. This pattern underscores a trade-off between ease of completion and the depth of data gathered.

Time Burden. Response times varied significantly across formats. The select words format was the quickest, requiring minimal cognitive and linguistic effort. The write text format was the most time-consuming, reflecting its open-ended nature and the additional time needed to construct coherent narratives. The write phrases and write words formats fell in between. These findings suggest that the choice of response format should consider both the intended depth of assessment and the practical constraints of the assessment context.

Limitations

While the study demonstrated strong validity and reliability of open-ended response formats, it remains limited by the use of cross-sectional survey data and rating scales, which are themselves imperfect indicators of psychological states. Individual differences in language ability or preference were not directly accounted for. Moreover, the order of question presentation may have introduced priming effects, and the models, while robust, were not fine-tuned for clinical language tasks. Future work should explore longitudinal designs, diverse populations, and personalized formats to further test generalizability and optimize language-based assessments.

Conclusions

Paper I showed that language-based assessments using both closed- and open-ended response formats can achieve moderate to strong psychometric properties for assessing depression and anxiety. Open-ended formats—particularly text and phrase responses—enabled richer, more nuanced expressions of mental health experiences compared to more constrained formats like select words. While even brief, structured language responses achieved high concurrent validity with traditional rating scales, more open-ended responses offered additional advantages in information richness, ecological validity, and external criterion validity (e.g., being associated with sick leave and healthcare use).

The findings also highlight important trade-offs: open-ended formats were more time-consuming for participants to complete and slightly less reliable across repeated assessments than structured formats. Additionally, while combining multiple formats improved assessing performance, it came at the cost of reduced discriminant validity between closely related constructs like depression and anxiety. Thus, while language-based assessments show promise as a complement to existing assessment tools, their implementation needs to carefully balance clinical depth, participant burden, and psychometric precision depending on the intended use. Paper I ultimately underscores the value of tailored response formats, chosen based on whether breadth, speed, or diagnostic specificity is prioritized.

Paper II

Understanding Suicidality and Self-Harm Through Probed Open-Ended Language: A Sequential Evaluation with Model Pre-registration

Aims

This study set out to develop, pre-register, and rigorously evaluate AI-based language models for assessing suicide risk and self-harm risk, using individuals' open-ended responses as the primary source of data. The models were trained to align with expert consensus ratings based on longitudinal, self-reported data collected over a 10-week follow-up period.

Participants were asked to respond to open-ended prompts covering general mental health, depression, anxiety, suicidality, and self-harm. The core hypothesis was that such language-based assessments would capture meaningful psychological markers that correspond with expert evaluations of suicidality risk and self-harm risk.

In addition to concurrent validity, the study examined the generalizability of the models through a formal pre-registration and held-out evaluation protocol using the sequential evaluation with model pre-registration framework. Ultimately, this work aimed to demonstrate that language-based models, grounded in individuals' own descriptions of their mental states and experiences, can serve as reliable and ecologically valid tools for suicide risk and self-harm assessment.

Methods

Participants. Participants were recruited online using Prolific. The dataset was collected in two phases as part of the sequential evaluation with model pre-registration design. The development sample consisted of 641 participants, and the pre-registered test sample included a new, held-out cohort of 150 individuals.

Procedure and Measures. As part of a larger study, participants completed a set of open-ended prompts targeting suicidality, depression, and anxiety, followed by validated rating scales and a comprehensive clinical history survey. The primary criterion for model evaluation was expert-rated suicidality and self-harm risk, based on longitudinal expert data (LED) reviews of repeated self-reports collected over a 10-week period. Models were trained on open-ended responses transformed into embeddings using large language models and evaluated with regularized regression. To ensure robust evaluation, the study followed a pre-registered two-phase sequential evaluation with model pre-registration framework: models were first trained and cross-validated on the development sample, then pre-registered and applied to an independent test sample. This design enabled a transparent assessment of model generalizability against an expert-derived reference standard.

Results

Concurrent Validity to Expert-Rated Assessments Using Held-Out Data with Pre-Registered Models. The pre-registered language-based assessment models demonstrated moderate to strong concurrent validity with expert-rated suicidality risk and self-harm in the held-out test sample. For suicidality, the model incorporating all language data reached a correlation of *r* = .70 (disattenuated *r* = .90), exceeding the pre-registered performance threshold. When using only responses to suicidality prompts, the model achieved *r* = .57 (disattenuated *r* = .73). Both models significantly outperformed the demographic-only model (*r* = .26, *p* < .001), underscoring the added assessing value of linguistically rich input. Similarly, for self-harm, the full-language model reached *r* = .68 (disattenuated *r* = .92), and the prompt-specific model reached *r* = .65 (disattenuated r = .89), both substantially outperforming the demographic baseline (*r* = .30, *p* < .001). These findings indicate promising validity of the language-based models as research tools in assessing complex mental health constructs, while further validation is required before considering clinical application.

Incremental Validity to Expert-Rated Assessments. The study also tested whether combining language responses from multiple domains—such as general mental health, suicidality, self-harm, anxiety, and depression—could enhance assessment performance. Results confirmed this hypothesis: integrating information across constructs yielded better convergence with expert assessments than models based on suicidality description. This highlights the multifaceted nature of suicide risk and suggests the potential utility of holistic modelling approaches for research purposes.

Discriminant Validity. The models exhibited meaningful discriminant validity, as their strongest correlations were with expert-rated suicidality risk—not with related but distinct constructs such as self-harm severity, depressive mood, or excessive worry. Although these constructs are often comorbid with suicidality, they are considered distinct phenomena with different underlying motivations. For example, self-harm is often characterized by non-lethal behaviors such as "cutting," "scratching," or "picking_skin," typically serving functions like emotional regulation or self-punishment, whereas suicidality involves ideation, planning, and intent to die—reflected in topics like "killing," "thought_killing," and "attempted_suicide." The lower correlations with non-targeted assessments demonstrate that the models captured construct-specific information rather than simply general emotional distress. These distinctions are consistent with DSM-5-TR (Diagnostic and Statistical Manual-5-Text Revised, American Psychiatric

Association, 2022) definitions of Suicidal Behavior Disorder and Nonsuicidal Self-Injury, and they illustrate the interpretability of language-based models, though their diagnostic value requires further evidence.

Face Validity (Development and Held-Out Samples). Visual analyses of the most predictive language features across both the training and held-out datasets provided evidence for face validity. Key terms such as "actived_upon," and "planned" were prominently associated with higher levels of suicidality risk, reflecting established clinical themes. These semantic patterns appeared theoretically coherent and may provide insight into how suicidal states manifest in language.

Limitations

Although expert-rated clinical ratings provided a strong reference standard, they were based on aggregated expert judgments rather than direct behavioral outcomes like hospitalization or suicide attempts. Additionally, the use of Swedish-language data and a relatively homogeneous sample may limit the generalizability of the findings across cultures or languages.

Conclusions

Paper II shows that language-based assessments, when grounded in individuals' open-ended descriptions of their mental states, can align well with expert ratings of suicidality and self-harm risk. By training models on prompted language and validating them against expert-rated risks rather than self-report scales, the study demonstrated that language-based assessments can approximate expert judgments with promising accuracy under research conditions.

The models showed consistent performance across development and held-out samples, supporting their robustness in a controlled research setting. Furthermore, the models differentiated suicidality risk from related constructs such as self-harm, indicating construct-specific associations rather than broad distress alone. Combining language data across multiple formats and construct domains improved performance, underscoring the value of a multidimensional approach for enhancing predictive alignment with expert assessments.

Visual analyses of predictive language highlighted that these models captured markers consistent with clinical theory—such as hopelessness, ideation, and emotional overwhelm—illustrating their interpretability and potential research utility, though their applied clinical value remains to be established.

Overall, this study supports the promise of language-based assessments as interpretable research tools for studying suicidality and self-harm risk. Their demonstrated ability to approximate expert ratings highlights their potential, while further validation against diagnostic interviews, behavioral outcomes, and real-world decision-making is needed before integration into clinical workflows.

Paper III

Language-Based Affect Assessments Capture Experiment-Induced Changes Beyond Rating Scales

Aims

The primary aim of Paper III was to examine the causal validity of language-based assessments of affect by testing their sensitivity to experimentally induced emotional changes under controlled conditions. The study compares language-based assessments and traditional closed-ended rating scales (specifically, the Positive and Negative Affect Schedule, PANAS) in their ability to classify affective responses following an experimentally administered mood induction procedure. In this between-subjects design, participants were randomly assigned to one of three distinct conditions designed to evoke different affective responses by being in: (1) a church, (2) a shopping mall, or (3) a natural park.

The assigned location in the mood induction procedure served as an external manipulation, providing a criterion against which to evaluate the sensitivity of affect measures to systematic changes in emotional states. If an assessment method is sensitive to actual changes in emotional states, we reasoned that it should show above-chance ability to classify participants into their assigned conditions based on their post-induction affect reports. This classification accuracy was interpreted as an indicator—rather than a direct proof—of the causal validity of each assessment method. To evaluate generalizability across settings, Paper III included both an online sample, where participants watched videos of the mood induction environments, and an offline sample, where participants physically visited the same filmed locations.

Methods

Participants. Participants were drawn from two separate samples to enable both model development and generalisability evaluations. The online sample was recruited through Prolific including 1000 participants for four conditions: church, mall, park and an open-air market. The main analyses focus on the three first conditions (see S5 for details regarding the open-air market). For the offline sample, 153 participants were recruited. The development sample consisted of 586 participants stratified by gender, age, and condition. Two independent evaluation samples were used to test generalizability: an online holdout sample ($N = 153$) and an offline sample (*N* = 153) recruited in Malmö in between the church, shopping mall and the park.

Procedure and Measures. The study used a between-subjects experimental design with random assignment to one of the three mood induction procedure locations. Participants were first asked to answer questions regarding their current affective states, personality and demographics questions. Then they were randomly assigned to experience one of three distinct locations: a church, a shopping mall, or a natural park. These conditions were selected to elicit different emotional responses through varying sensory and social contexts. Offline participants physically visited one of the locations, whereas online participants watched a 201-second video of the same environments, with instructions to be mindful and "take in the environment". Following the induction, participants described their emotional experience using an open-ended prompt ("Please describe how you felt in the location") and also completed the Positive and Negative Affect Schedule.

Results

The language-based assessments showed higher accuracy than Positive and Negative Affect Schedule in detecting experimentally induced changes in affect, across training, online, and offline samples.

Classification Accuracy Across Datasets. The language-based assessment generally achieved stronger performance in detecting experimentally induced changes in affect over the traditional Positive and Negative Affect Schedule rating scale in classifying participants' assigned mood induction conditions (church, shopping mall, or park). In the cross-validated training dataset, language-based assessment achieved a significant difference in area-under-the-curve of .67 -.74, compared to .39 - .72 for the Positive and Negative Affect Schedule. This significant performance advantage persisted in both prospective evaluations: in the online holdout sample, language-based assessment achieved an area-under-the-curve of .72 - .74 versus .58 - .72 for the Positive and Negative Affect Schedule; in the offline holdout sample, language-based assessment maintained an area-under-the-curve of .67 - .69, exceeding the Positive and Negative Affect Schedule area-under-the-curve of .39 - .53. While some comparisons were statistically significant, differences between language-based assessments and Positive and Negative Affect Schedule were modest in the online sample and clearer in the offline sample. Pre-trained language-based assessments of valence provided clearer differentiation between conditions than Positive and Negative Affect Schedule, with larger between-condition effects (F = 34.65, η^2 = .086) compared to Positive and Negative Affect Schedule (F = 9.37, η^2 = .025). In addition, discrete emotions such as anger showed higher classification accuracy than PANAS in both the online (area-under-the-curve = .66) and offline (area-under-the-curve = .64) samples, while other emotions including joy, trust, disgust, and fear reached area-under-the-curves of .58–.65 in at least one dataset where Positive and Negative Affect Schedule did not perform significantly. Taken together, these theoretically grounded models showed at least comparable, and in some cases stronger, sensitivity than Positive and Negative Affect Schedule in capturing condition-related affective differences.

Face Validity and Descriptive Richness. Language visualizations further provided evidence for face validity, illustrating that language-based assessments captured condition-specific emotional expressions. The open-ended responses demonstrate that language-based assessments can offer complementary descriptive insights of the affective state related to each experimental setting. These patterns underscored the ecological and interpretive value of language-based data, offering insight not only into classification accuracy but also into the nature of the affective experiences being reported.

Limitations

The mood induction study prioritized ecological validity by using immersive environments (e.g., churches, parks), yet these naturalistic settings may not elicit consistent emotional responses across individuals or samples. Moreover, while language-based assessments demonstrated strong generalizability, the corresponding Positive and Negative Affect Schedule model failed to replicate in the offline sample, suggesting limited robustness of traditional scales under ecologically complex conditions. The use of classification accuracy as a proxy for causal validity also provides only a partial picture of participants' nuanced affective states. Future work should consider more targeted emotion inductions and further explore how traditional rating scales compare to language-based models in detecting subtle or contextually driven shifts in affect.

Conclusions

Paper III provides evidence that language-based assessments are not only capable of detecting affective changes resulting from experimental manipulations but do so significantly more effectively than traditional rating scales. Across multiple datasets and testing conditions, language-based assessments generally outperformed the Positive and Negative Affect Schedule scale in classifying participants' assigned mood induction conditions, underscoring their sensitivity to real-time emotional shifts.

In addition to their quantitative accuracy, language-based assessments offered richer and more nuanced insights into participants' emotional states, as demonstrated through word-level visualizations that captured meaningful variations in affective expression across experimental settings. These qualitative advantages reflect the expressive depth of open-ended language, which appeared especially well-suited to capturing subtle emotional distinctions.

The robustness of language-based assessments was further indicated by their consistent performance across levels of measurement granularity, whether the Positive and Negative Affect Schedule was analyzed as a single composite or as individual items. Together, these findings provide preliminary evidence for the causal validity of language-based assessments and highlight their potential as flexible, expressive, and empirically grounded tools for assessing affective experiences in experimental contexts.

Paper IV

The language-based assessment Model (L-BAM) Library: Open Model Sharing for Independent Validation and Broader Applications

Aims

The aim of Paper IV was to increase the accessibility, reproducibility, and broader adoption of language-based assessment models by introducing a centralized and standardized framework for their application and dissemination: the language-based assessment Models (L-BAM) Library. This open-access resource was developed to support social and psychological scientists in discovering, evaluating, and applying pre-trained L-BAMs for a range of psychological constructs while also encouraging independent validation and model sharing within the research community.

To achieve this, the paper presents two key tools. First, it introduces the *textAssess()* function, which enables users to automatically download models in L-BAM, preprocess language data, and apply models for prediction, classification, or psychological assessment—lowering the technical barrier for researchers. Although this tool can also be used by practitioners, the paper emphasizes that rigorous evaluation of model suitability remains essential before any applied use. Second, it provides an overview of the L-BAM Library, an online repository where users can explore existing models and contribute new ones by following clear documentation standards for usage, citation, and metadata.

Methods

Paper IV introduced the L-BAM Library and accompanying tools to support open, reproducible, and scalable language-based psychological assessment. The core analyses and infrastructure in this paper were built using the text R package, which includes functionality for model deployment, text preprocessing, embedding generation, and prediction. These tools rely on back-end integration with the reticulate package to bridge R and Python, and the transformers Python library for state-of-the-art embedding models.

All analyses are reproducible and compatible with open science workflows. Example code was provided in the paper through boxed demonstrations for assessing depression, suicidality, and experimental affective condition using downloadable models from the L-BAM Library.

Contributions

Paper IV makes a methodological contribution to the field of psychological assessment by introducing the L-BAM Library—an open-access resource for sharing, applying, and building upon language-based assessment models. The library currently hosts a broad collection of pre-trained L-BAMs targeting a diverse range of psychological constructs, including mental health variables (e.g., depression, anxiety; Gu et al., 2025), well-being (e.g., life satisfaction, harmony in life, Kjell et al., 2022; autonomy, Mesquite et al., 2025), personality (e.g., implicit motives; Nilsson et al., 2025), and expert-rated risk assessments (e.g., suicidality risk, self-harm risk, Gu et al., 2025).

To enhance accessibility and encourage broad usage, the paper also introduces the *textAssess()* function—a tool designed to streamline the application of L-BAMs. This function allows users to automatically download pre-trained models, pre-process open-ended language responses, and generate psychological assessments or classifications with minimal technical overhead. The tool is designed with both researchers and applied professionals in mind, lowering the barrier to entry for incorporating computational language assessments into psychological research and practice.

In addition to providing this infrastructure, the paper presents a detailed tutorial on how to use both the L-BAM Library and the text package. This includes step-by-step guidance on how to search for existing models, and contribute new models that adhere to standardized formatting and documentation practices. The tutorial emphasizes transparency and reproducibility, supporting open science efforts in the domain of natural language-based psychological assessment.

Together, these tools are intended to streamline the application of L-BAMs, promote methodological transparency, and foster open collaboration in the field of language-based psychological assessment. By making models accessible and encouraging their careful use, Paper IV seeks to support open and cumulative research in psychology and related fields, while emphasizing the need for independent validation and cautious application. To illustrate the practical implementation of L-BAMs, below I show how models from Paper I, II and III can be applied:

Code box 1 - Example on depression severity

```
# Example text to access
text_to_assess = c(
   "I feel down and blue all the time.",
"I feel great and have no worries that bother me.")

library(text) # see Code box 2 if the package has not been installed

# Predict depression severity scores using a model trained with the text package
# Download the model, create word embeddings, and apply the model to these word embeddings.
depression_scores <- textAssess(
model_info = "depression_text_phq9_roberta23_gu2024",
texts = text_to_assess,
dim_names = FALSE)

# Output
depression_scores


# A tibble: 2 × 1
  `word_embeddings__PHQ9$PHQtotpred`
                             <dbl>
1                             17.2
2                             4.10
```

Code box 2 - Install the text-package

```
# Step 1: Install the text-package in R
install.packages("text")
library(text)

# Step 2: Set up the required Python environment
# The function first checks that common system dependencies are satisfied and if not provide
instructions for how to install them in the terminal.
textrpp_install()

# Step 3: Initialize the Python environment for use with text
textrpp_initialize()

# If you encounter any issues during installation or setup, please refer to the latest instructions and
troubleshooting guide at:
https://www.r-text.org/articles/ext_install_guide.html
```

Code box 3 - Example on Valence and Well-being

```
# Assess the valence of the harmony in life texts
# Download the model, create word embeddings, and apply the model to these word embeddings.
valence_scores <- textAssess(
model_info = "valence_facebook_mxbai23_eijsbroek2024",
texts = Language_based_assessment_data_8$satisfactiontexts)

# Correlate the assessed valence scores with the harmony in life scores
cor(valence_scores$texts__Valencepred,
  Language_based_assessment_data_8$swlstotal)

[1] 0.7421613
```

Limitations

While Paper IV advances open science through the creation of the L-BAM Library, the current set of models remains limited, and long-term standardization depends on user engagement and sustained documentation quality. Further, the generalizability of language-based models across contexts is not guaranteed. Since model performance can vary depending on the setting, population, evaluation contexts, and language distribution, users must critically evaluate whether a given model is appropriate for their specific use case. The library therefore emphasizes transparent reporting of model training data and performance to support responsible and context-sensitive applications. In addition, users must remain attentive to ethical concerns such as privacy, transparency, and accountability, especially when models are applied in sensitive or clinical domains.

Conclusions

Paper IV introduces a comprehensive infrastructure to support the transparent, accessible, and scientifically rigorous application of L-BAM in psychological science. By launching the L-BAM Library alongside the text R package and its core function *textAssess()*, this paper addresses key barriers to adoption, including technical complexity, lack of standardized resources, and limited support for model sharing and validation.

The L-BAM Library consolidates a growing collection of pre-trained models covering a wide array of constructs offering researchers a centralized platform for discovering, applying, and contributing models. Many of these models have demonstrated high convergent and criterion validity.

Together, these contributions aim to advance the role of language as a core measurement modality in psychology. By offering user-friendly tools, standardization protocols, and clear pathways for model contribution and reuse, Paper IV provides initial groundwork for a more cumulative, collaborative, and resource-efficient approach to computational psychological assessment, while emphasizing the need for cautious interpretation and ongoing validation rather than claiming to fully bridge research and applied practice.

Chapter 6. Ethical Considerations and AI Safety

Introduction

Ethical responsibility is a foundational pillar of both psychological research and the development of AI. As this thesis introduces new methods for assessing mental health using AI-driven language-based tools, it becomes essential to examine ethical considerations from two interconnected perspectives: those governing research with human participants, and those specific to the design, implementation, and application of AI technologies.

From a research ethics standpoint, studies involving mental health assessments demand heightened sensitivity to issues of informed consent, participant autonomy, privacy, and potential psychological harm. These responsibilities are amplified when open-ended language data is collected, as such responses may include personal, emotionally charged, or identifiable content. Ethical research design must therefore ensure that participants' rights are protected, that data collection is transparent, and that consent procedures are robust and contextually appropriate (Jobin et al., 2019; Kjell et al., 2024; Kurita et al., 2019; Lison et al., 2021; Leidner, & Plachouras, 2017; Shah et al., 2020).

At the same time, the increasing use of AI in psychological assessment raises a new set of ethical questions—about fairness, interpretability, accountability, and the potential for unintended consequences. AI systems, particularly those based on large language models, are not neutral tools: they reflect the data on which they are trained, and their outputs can carry significant implications for individuals' mental health evaluation and treatment. As such, ensuring transparency, minimizing bias, and fostering clinician and user trust are critical for responsible deployment (Krieger et al., 2024; Lawrence et al., 2024; Timmons et al., 2023).

This chapter explores these dual dimensions of ethics. It begins by outlining the core principles of research ethics relevant to the studies conducted in this thesis. It then considers broader ethical frameworks for AI, including recent guidelines and normative frameworks for trustworthy AI. Together, these perspectives inform the ethical foundation upon which the models and methodologies in this thesis were built, and guide their future use.

Ethical Considerations in Current Research

The research presented in this thesis was conducted in alignment with fundamental ethical principles of psychological science. Each of the included studies received approval from the Swedish Ethical Review Authority, ensuring that data collection and participant involvement adhered to national and international ethical standards (Swedish Ethics Application 2020-00730).

To promote scientific transparency and minimize analytical flexibility, assessment models were pre-registered prior to evaluation. This use of pre-registration helped safeguard against researcher degrees of freedom and ensured a confirmatory approach to model testing, reinforcing both the ethical and methodological integrity of the work.

Ethical Considerations in Suicide research

Previous research showeds that asking people about their mental health and suicidal ideation, even when intensively or repeatedly, did not trigger suicidal or self-harm behavior and did not increase suicidal ideation (Bender et al., 2019; Cukrowicz et al., 2010; DeCou & Schumann, 2018; Gould et al., 2005; Hom et al., 2018; Mathias et al., 2012; Kivelä et al., 2024). Further, Dazzi, et al. (2014) found that "Recurring ethical concerns about asking about suicidality could be relaxed to encourage and improve research into suicidal ideation and related behaviors without negatively affecting the well-being of participants." They pointed out that "there is a commonly held perception in psychology that enquiring about suicidality, either in research or clinical settings, can increase suicidal tendencies." A review of the published literature examining whether enquiring about suicide induces suicidal ideation in adults and adolescents, and general and at-risk populations showed that: "None found a statistically significant increase in suicidal ideation among participants asked about suicidal thoughts." Their findings "suggest acknowledging and talking about suicide may, in fact, reduce, rather than increase suicidal ideation, and may lead to improvements in mental health in treatment-seeking populations." (Dazzi et al., 2014).

Informed Consent and Participant Rights

Participants were fully informed about the purpose, procedures, and nature of each study through clearly written consent forms. These forms emphasized that participation was entirely voluntary and that individuals could withdraw at any time without any consequence. Given the sensitive nature of the topics explored—such as depression, anxiety, and suicidality—special care was taken to communicate that all responses were collected anonymously and that no individual-level data would be monitored or followed up. This transparency supported informed decision-making and upheld participants' rights to privacy and autonomy throughout the research process.

Handling Sensitive Mental Health Data

Given the highly personal and potentially distressing content of open-ended responses related to depression, suicidality, and anxiety, the studies implemented safeguards to ensure ethical data handling. Participants were explicitly informed that their responses would not be monitored in real time and would not result in any clinical intervention. To reinforce this understanding, a mandatory checkbox was included after answering questions about suicidal ideation, confirming that individuals acknowledged the anonymous nature of the data collection and the absence of researcher follow-up.

Participant Support and Resources

To support participant well-being, studies on mental health prominently displayed suicide prevention and mental health helpline information both before, during (were questions about suicide were presented) and after participation. In more emotionally sensitive contexts in Paper III—such as mood induction procedure—participants were screened in advance using brief validated tools (i.e., the Patient Health Questionnaire-2, PHQ-2, Arroll et al., 2010). Those indicating elevated symptoms were excluded from participation to minimize the risk of distress during the study. These practices ensured a protective ethical framework for engaging with vulnerable populations.

Privacy and Re-identification Risks

The use of open-ended language responses empowered participants to articulate their psychological experiences in their own words, rather than being limited to pre-defined categories. This design reflects a commitment to respecting participant autonomy and expression, allowing for more personalized, nuanced, and meaningful engagement with the assessment process. However, open-ended language data presents unique challenges for participant privacy. While we did not explicitly collect personally identifiable information such as names, addresses or IP-addresses, the free-text nature of these data inherently increases the risk of re-identification. Participants may unintentionally disclose information that could be used to infer their identity, especially when responses include idiosyncratic details, rare experiences, or identifiable phrases.

To reduce the potential for re-identification, several precautionary measures were implemented throughout the research process. First, a principle of data minimization was followed—only the information essential for analysis was collected, and responses were stripped of any explicit identifiers before processing. All data were securely stored on encrypted servers, with access restricted to authorized research personnel. Additionally, pseudonymization techniques were applied, ensuring that any links between participants and their responses were removed or replaced with non-identifying codes. These safeguards were designed not only to comply with regulatory standards such as the GDPR but also to uphold participants' trust when sharing sensitive personal experiences in their own words.

Predictive Inference and Ethical Risks

Following the broad overview of ethical principles, it is critical to consider specific risks that arise from the use of assessment technologies in health and psychological contexts. Especially with AI models that can infer sensitive, un-volunteered information from language, new ethical challenges emerge that go beyond classical concerns about privacy and consent (Siegel, 2020). The following discussion highlights key issues that are particularly relevant to language-based mental health assessments.

Sensitive Inferences Without Consent

One major ethical risk is that assessment models can infer extremely sensitive personal information—such as pregnancy status, serious health conditions, sexual orientation, or even the likelihood of imminent death (Mühlhoff, 2023) —even if individuals have not explicitly disclosed this information. In the context of language-based mental health assessments, this risk is amplified. Open-ended language responses, by their very nature, may reveal psychological vulnerabilities, trauma histories, or other highly sensitive attributes that extend beyond the intended scope of the original assessment. Such inferences, although technically powerful, raise serious concerns about participant autonomy and informed consent, as individuals may not anticipate the full range of personal information that could be extracted from their language.

Risks of Misuse and Discrimination

A second major concern involves the potential misuse of assessed outputs. Information inferred from language could be used for profiling, surveillance, exclusion, or discriminatory practices. In the mental health domain, mismanaged inferences about conditions like depression, anxiety, or suicidality could result in harmful stigmatization, affect access to services, or unjustly influence decisions made by employers, insurers, or legal authorities. Without strong ethical controls, there is a risk that AI-driven assessments could inadvertently reinforce biases or create new forms of inequity, particularly if the models are applied in contexts beyond those originally intended.

Responsibility for Indirect Data Generation

Assessment technologies also differ from traditional data collection methods in that they may generate "new" data by inferring from seemingly innocuous information. This creates a novel ethical burden: even if individuals have not directly provided certain information, AI systems can effectively "manufacture" sensitive data from patterns in their language. Researchers and developers thus bear a heightened responsibility to recognize that the act of assessment itself can constitute a privacy and autonomy risk. It is not enough to protect only the raw inputs; the assessed outputs, especially those touching on deeply personal matters, must be treated with similar caution and ethical consideration.

Mitigation Strategies

Several mitigation strategies are necessary to address these emerging risks. First, clear boundaries should be established regarding what types of inferences are ethically permissible from language data, ideally with participant consent covering the possibility of indirect inferences. Second, researchers must communicate transparently with participants about what the models may assess and the limits of confidentiality. Participants should be made aware that even seemingly neutral responses might lead to sensitive assessments. Third, strong governance frameworks should be put in place to manage the storage, sharing, and interpretation of sensitive assessed outputs. Access to model assessments should be carefully limited, and interpretations should be made with humility, acknowledging the probabilistic nature of AI models. Aligning assessing efforts with the ultimate goal of participant welfare, rather than mere assessment accuracy or commercial gain, is critical for responsible innovation in psychological assessment.

Ethical Considerations in Applied Settings and Core AI Principles

As AI-driven psychological assessments move from research into applied settings, especially within clinical or commercial domains, ethical considerations become more complex. The challenges are not limited to the accuracy or validity of these tools, but also involve broader concerns such as responsible governance of assessment technologies. These concerns are particularly heightened when AI systems are developed or deployed by private companies, where incentives may differ from academic or clinical priorities. To guide the ethical development and deployment of language-based AI tools, it is essential to draw on established ethical frameworks. In recent years, a growing body of interdisciplinary research has identified core principles that should govern the responsible use of AI in health and social contexts. Two major reviews—Jobin et al. (2019), which synthesized over 80

AI ethics guidelines worldwide, and Corrêa et al. (2023), have helped distill a set of foundational principles. These include transparency, justice, non-maleficence, responsibility, privacy, beneficence, autonomy, trust, sustainability, dignity, and solidarity. Together, they offer a robust ethical compass for designing and deploying AI systems in a way that upholds human values and social responsibility. Both reviews emphasize that the topic on AI ethics and safety is evolving quickly.

Further, as language-based AI tools are increasingly discussed for potential real-world deployment, especially in clinical or healthcare-adjacent settings, it remains crucial to emphasize that such deployment is still preliminary and must be preceded by comprehensive clinical trials and real-world evaluations where, regulatory frameworks are becoming central to ensuring their safe, transparent, and accountable use. Within the European context, two prominent regulatory instruments are particularly relevant: the General Data Protection Regulation (GDPR) and the recently passed EU AI Act. The GDPR sets out comprehensive rules for data protection, including strict requirements for informed consent, data minimization, and transparency in how personal data are processed. These principles are especially pertinent to language-based assessments, which can involve highly sensitive personal disclosures in open-ended formats. In addition, the EU AI Act introduces a tiered, risk-based framework for AI regulation. Under this law, AI systems used for mental health assessment—especially those influencing medical decision-making—are typically categorized as "high-risk" applications, underscoring the importance of rigorous validation and oversight before clinical deployment. According to Article 6 and Annex III of the AI Act (Regulation [EU] 2024/1689), such systems must meet stringent requirements, including comprehensive documentation, explainability, human oversight, and post-deployment monitoring. These legal expectations closely align with established clinical ethics around informed consent, diagnostic accountability, and the duty to avoid harm, and reinforce the need to treat current models as exploratory until further clinical validation is achieved.

While a full legal and technical analysis of these regulatory frameworks is beyond the scope of this thesis, their ethical implications underscore the importance of embedding safety, fairness, and accountability into all stages of AI assessment development. As the regulatory landscape continues to evolve, close alignment between psychological science and AI governance will be essential for translating research innovations into ethically sound clinical applications.

AI assessments in practice

While legal compliance and harm prevention form the bedrock of responsible AI deployment, ethical reflection must extend beyond regulatory boundaries. The use of computational models to assess psychological states introduces more fundamental questions about the nature of understanding, power, and care in mental health contexts. As Mittelstadt (2016) argues, algorithmic systems often introduce opaque forms of decision-making authority, raising critical concerns about accountability, interpretability, and value alignment in sensitive domains such as health care. Echoing this concern in a mental health context, Warrier et al. (2023) emphasize that AI tools must not be evaluated solely by their technical efficacy but also by how they affect the therapeutic relationship, diagnostic framing, and individual dignity.

Language-based AI systems may formalize and quantify aspects of human experience that have historically been understood through relational, narrative, or phenomenological frameworks. This raises the concern that subjective distress might be reduced to probabilistic output, undermining the inherently interpretive nature of clinical care. If a model flags someone as "high risk" for suicidality based on linguistic markers, does that designation gain authority over the person's own narrative? What are the implications of acting—or not acting—on such predictions?

Moreover, the epistemic authority of AI systems introduces new power asymmetries. Clinicians may feel pressure to defer to model outputs, while patients may perceive their experiences as being assessed impersonally. This dynamic risks displacing human empathy with technical precision. In worst-case scenarios, AI assessments could be used not to support care, but to justify exclusion from insurance, employment, or treatment, particularly among marginalized groups. Mittelstadt (2016) warns that without ethical safeguards, algorithmic systems may entrench existing institutional logics and reproduce social inequalities under the guise of objectivity. Warrier et al. (2023) similarly caution that when AI systems lack transparency and contextual nuance, they may exacerbate stigma, misclassification, or over-pathologization—especially among vulnerable or culturally diverse populations. Finally, there is an ontological question: What vision of mental health do these models encode? AI models trained on historical data risk perpetuating dominant clinical norms and cultural assumptions. The ethical stakes thus include not just fairness or privacy, but the definition of psychological normality itself. As highlighted by Mittelstadt (2016), algorithms not only reflect but also shape the normative frameworks through which human behavior is judged. As such, responsible innovation requires a participatory approach that includes clinicians, ethicists, patients, and affected communities in shaping how these tools are developed and deployed. Warrier et al. (2023) advocate for such participatory frameworks, stressing the importance of co-design with stakeholders to ensure ethical alignment with the lived realities of those affected.

Summary

This chapter has outlined the ethical foundations and safety considerations underlying the development and deployment of AI-driven language-based assessments for mental health. Ethical principles such as respect for autonomy, informed consent, and the protection of participant privacy were central throughout the empirical studies. Specific safeguards were implemented when collecting sensitive language data—particularly in relation to depression and suicidality—including anonymous data handling, clear disclaimers regarding the absence of clinical monitoring, and the provision of mental health resources.

In parallel, this chapter engaged with broader ethical concerns raised by the use of AI in psychological contexts. It discussed how AI-specific risks—such as data re-identification, —were managed through practices such as data minimization, secure storage, and model transparency. Drawing from leading ethical frameworks reviews (Jobin et al., 2019; Corrêa et al., 2023), the chapter synthesized core principles—such as transparency, justice, non-maleficence, and beneficence—that should guide responsible AI use in mental health assessment.

Chapter 7. General Discussion

Overview of Key Findings

This thesis sets out to develop and evaluate AI-driven language-based assessments for measuring key psychological constructs such as depression, anxiety, affective states, suicidality risk, and self-harm risk. Grounded in theoretical and psychometric frameworks (Kjell et al., in progress), the work demonstrates that natural language—when processed using techniques from natural language processing and large language models —can provide valid, reliable, and interpretable indicators of psychological states. Across four papers, the thesis addresses its core objectives: to construct models that translate open-ended language into quantitative mental health indicators; to evaluate their psychometric properties, including multiple forms of validity and reliability; and to foster open science by pre-registering evaluation strategies and developing reusable model libraries for broader application and transparency.

Paper I presented the first systematic evaluation of various natural language response formats in language-based assessments, ranging from structured formats like word selection to more open-ended formats such as writing phrases and full texts. The findings demonstrated that all formats achieved moderate to strong concurrent validity with established rating scales, and the more open-ended formats provided richer linguistic information and stronger convergence with external indicators such as self-reported sick leave and health-care visits. Notably, combining multiple formats led to increased assessment accuracy, underscoring the value of capturing diverse modes of expression.Paper I presented the first systematic evaluation of various natural language response formats in language-based assessments, ranging from structured formats like word selection to more open-ended formats such as writing phrases and full texts. The findings demonstrated that all formats achieved strong concurrent validity with established rating scales, but the more open-ended formats provided richer linguistic information and stronger convergence with external indicators like self-reported health-care outcomes. Notably, combining multiple formats led to increased assessment accuracy, underscoring the value of capturing diverse modes of expression. These results highlight that different formats contribute complementary insights, and that even brief, word-based formats can be highly informative when combined with more expressive language responses—offering flexibility for both time-sensitive and in-depth clinical and research contexts.

Paper II extended this work by developing and evaluating language-based models of suicidality and self-harm, benchmarked against expert-rated risk assessments. The study showed that language-based assessments align with expert judgments in a held-out, pre-registered evaluation and generalized across independent samples, and added value by providing language descriptions. These results support the promise of language-based assessments as research tools, while further validation against diagnostic interviews, behavioral outcomes, and real-world decision processes is needed before any applied clinical use is considered.

Paper III demonstrated that language-based assessments are not only correlationally valid but also sensitive to experimentally induced changes in affective states. Compared to traditional self-report measures, language-based assessments more accurately classified participants into mood induction conditions, showing that language-based assessments can detect shifts under controlled experimental conditions.

Finally, Paper IV addressed the challenge of accessibility by introducing the L-BAM Library and supporting software tools. This infrastructure enables open dissemination, reproducibility, and collaborative development of language-based assessments, lowering barriers primarily for researchers to apply and independently validate these tools in health-related research. Potential applied use requires careful, context-specific evaluation and additional external validation.

The progression of validation strategies across the thesis can be understood as a gradual effort to strengthen the evidentiary basis for the proposed models—moving from subjective self-report, to expert-derived assessments, and finally to experimentally manipulated conditions. In Paper I, language-based assessments are examined in relation to traditional rating scales, with the aim of ensuring comparability with widely used self-report instruments. Paper II extends this by comparing model outputs to expert ratings derived from longitudinal patient data, providing a more contextually grounded reference point. Paper III develops the approach further by assessing the models' sensitivity to experimentally induced changes in affective states through mood induction procedures. This sequence—from self-report to expert report to experimental manipulation—suggests a progression toward stronger forms of evidence, while also indicating the potential adaptability of language-based models across different sources of psychological data.

The thesis engages with the current landscape of psychometric assessment by considering how language-based assessments might both align with established methodologies, such as rating scales, and extend beyond them. Rather than adhering strictly to a single validation tradition, the work explores how multiple approaches—self-report, expert judgment, and experimental manipulation—may be brought together in ways that invite more flexible perspectives within the field. For example, Paper III illustrates how language-based assessments can be both theoretically informed and data-driven: the models are grounded in meaningful constructs, yet remain responsive to experimentally induced changes, and are able to produce descriptive insights through data-driven visualizations. In this way, the thesis suggests that language-based models may hold value across explanatory, predictive, and descriptive domains, offering possibilities for more nuanced and context-sensitive approaches to psychological assessment.

However, future research will need to more fully realize the transformative potential of language-based assessments by moving beyond the boundaries of existing diagnostic frameworks. While the current thesis demonstrates that language-based assessments can validly and reliably assess constructs like depression, anxiety, and suicidality, these constructs are still treated as stable targets—implicitly reinforcing the assumption that such categories exist independently of the tools used to define and measure them. To unlock the full potential of natural language, future work should explore how language-based assessments can challenge, refine, or even replace conventional psychiatric categories by uncovering the heterogeneity, ambiguity, and socially constructed dimensions of psychological suffering. This includes examining and evaluating language-based assessments with dimensional models such as HiTOP (Hierarchical Taxonomy of Psychopathology; Kotov et al., 2017), which conceptualize mental health as continua rather than discrete diagnoses, and investigating how language-based assessments can detect meaningful psychological change in clinical interventions beyond symptom reduction alone (e.g., see Ekstrand, 2024). In doing so, language-based assessments may help reframe what counts as improvement, recovery, or even disorder—supporting more individualized, contextualized, and theoretically expansive approaches to psychological science and practice.

Integration with previous research

The results of this thesis extend, support, and in some cases nuance existing psychological and computational research on the use of contextualized language for assessing mental health. Language-based assessments have gained increasing prominence in psychological science, particularly with the growing interest in using natural language as both an expressive and measurable reflection of mental states. The studies in this thesis contribute to this expanding field by demonstrating the feasibility, reliability, and validity of probed language-based assessments—responses collected via questions. To situate these findings, it is essential to contrast them with prior work based on contextualized language—text data that arises naturally in the course of everyday communication, including social media posts on

Reddit, Facebook, Twitter, as well as clinical information such as Electronic Health Records, and clinical interviews.

Research Using Social Media posts

A large body of research has investigated the use of naturally occurring language from social media platforms to infer mental health status. Reddit has emerged as a prominent source in this domain due to its anonymity and topic-specific subreddits, such as r/depression or r/SuicideWatch. Studies have shown that language in user posts can be predictive of psychological conditions including depression and suicidality (e.g., Gkotsis et al., 2017; Ji et al., 2021; Low et al., 2020; Xu et al., 2024; Yates et al., 2017). These models are often trained on data from individuals self-identifying with mental health struggles, offering insight into real-world linguistic patterns of distress. However, such studies often suffer from limitations like the absence of clinical ground truth labels. Furthermore, these methods do not involve standardized questions, making it difficult to ensure construct alignment.

In parallel, Facebook-based research has demonstrated that user-generated posts can assess depression and emotion-related constructs (Eichstaedt et al., 2018; Katchapakirin et al., 2018; Schwartz et al., 2014). These studies benefit from longitudinal data and allow analysis of personal trajectories and behavior patterns over time. However, like Reddit research, they often lack direct validation against clinical measures (although see Eichstaedt et al., 2018 for an exception) and offer limited interpretability due to the absence of directly probed questions. In contrast, the current thesis explicitly probes mental health constructs through standardized open-ended questions and validates outcomes using concurrent measures, thus enabling higher convergence with corresponding rating scales.

Twitter (now X.com) has also been extensively studied, particularly for its potential in early detection of mental health crises such as suicidality (e.g., Coppersmith et al., 2018; De Choudhury et al., 2016), and affective disorders (Coppersmith et al., 2014, & 2015). The platform's time-stamped, public micro-posts offer unique opportunities for real-time monitoring and temporal modeling. However, the short length of tweets restricts expressive depth, and the informal, often sarcastic or performative style of language on Twitter poses challenges for accurate mental health assessment. In contrast, the elicited language in this thesis enables more detailed and descriptive accounts of psychological states, offering a higher degree of face validity and interpretability than what is typically available through tweet analysis. Beyond social media platforms, smartphone-based behavioral sensing has also been used to predict personality (Stachl et al., 2020), highlighting the ecological value of digital traces as complementary to language.

Research Using Clinical Information

Research using Electronic Health Records (EHRs) has become increasingly common in computational psychiatry, particularly in efforts to extract mental health indicators from clinical notes. Such studies often rely on language produced in healthcare settings and annotated by professionals, making them highly relevant for applied clinical contexts (e.g., Adams et al., 2024; Cardamone et al., 2025; Guevara et al., 2024; Shickel et al., 2017). However, electronic health records data come with several limitations: they are typically not publicly accessible due to strict privacy constraints, and the text content reflects clinician interpretations rather than the patient's own words. This introduces a layer of filtering that may obscure the subjective experience of mental health. In contrast, the language-based assessments developed in this thesis directly elicit responses from individuals, offering unmediated access to their lived experiences. This enhances ecological validity and enables a more person-centered form of psychological measurement.

Language derived from clinical interviews, whether structured or semi-structured, has also been used in research to uncover linguistic patterns associated with psychological conditions (e.g., Cohen et al., 2020; Tanana et al., 2021; Wright-Berryman et al., 2023). These interviews provide rich contextual data and benefit from professional engagement, allowing for in-depth exploration of mental health phenomena. However, they are often resource-intensive, requiring trained interviewers and significant time to conduct and transcribe. Moreover, interviewer presence may shape participant responses, leading to potential bias or variability that limits generalizability. The approach taken in this thesis offers an alternative: standardized, self-administered language prompts that remove the need for an interviewer while maintaining consistency across participants. This structure preserves the depth of open-ended expression while supporting reliable comparisons across different individuals and samples. However, a limitation of probed language-based assessments is that clinicians in real-world settings may ask novel or spontaneous questions that fall outside the predefined prompts. As a result, future work should explore ways to make language-based assessments more flexible and adaptable—allowing for dynamic questioning while preserving standardization and comparability.

Integrating Language and Context in Psychological Assessment

While this thesis focuses on language-based assessments derived from elicited, open-ended responses, it does not directly compare their assessment accuracy to models based on social media language or electronic health records. These different sources of language are not mutually exclusive and may in fact complement one another—each offering distinct insights depending on the context of use (as noted in Chapter 1; Bhatia, & Aka, 2022; Boyd, & Schwartz, 2021; De Choudhury et al., 2013; Demszky et al., 2023; Dumas et al., 2025; Eichstaedt et al., 2015; Kjell et al., 2023b, & 2024; Sametoğlu et al., 2024). Moreover, language-based assessments can be meaningfully combined with traditional rating scales to enhance precision, interpretability, and coverage of psychological constructs. Future research should explore how integrating multiple data sources (Singh et al., 2025) may yield more robust and context-sensitive assessments (e.g., Varadarajan et al., 2024).

In line with prior studies demonstrating using natural language for psychological constructs via adding probed questions (e.g., Kjell et al., 2019, & 2022), this work reinforces that both closed- and open-ended language-based response formats can serve as valid and reliable tools for assessing depression, anxiety, and related affective states. Paper I provides a systematic investigation of multiple language response formats and their relationship with established self-report scales, illustrating that even minimal open language (e.g., a few words or phrases) can capture psychological variance typically measured by multi-item rating scales.

Notably, Paper II bridges the gap between language-based assessment methods and expert assessments, a topic of increasing interest in clinical AI research. However it was not possible to compare language-based assessments directly to rating scales—since the expert clinical scores themselves were based on both language responses and numerical scales.

Moreover, Paper III advances the discourse on assessment and causal validity, demonstrating that language-based assessments are sensitive to subtle emotional changes induced through experimental manipulation, even outperforming rating scales. This supports the position that open-ended responses may better capture the richness of affective states, corroborating earlier work on expressive writing and digital affective monitoring (Pennebaker & Beall, 1986; Bakker & Rickard, 2018), while offering a scalable AI-based method for real-time affect tracking.

Across all papers, the findings integrate with and support ongoing research into person-centered, context-sensitive, and interpretively rich assessments of mental health, as advocated by contemporary psychological theory (Boyd & Markowitz, 2025; Kjell et al., 2024). This thesis therefore positions itself at the intersection of clinical psychology, computational psychiatry, and ethical AI, contributing to a cumulative science of mental health assessment that is both empirically grounded and theoretically informed.

Comparative Evaluation of language-based assessments

The studies in this thesis collectively position language-based assessments as a scientifically rigorous and practically valuable complement to traditional psychological assessment methods. Compared to closed-ended rating scales and structured interviews, language-based assessments demonstrate promising performance across core psychometric criteria under research conditions. It is difficult to compare language-based assessment and rating scales, because there is no objective truth of psychological constructs—mental health symptoms like depression or anxiety are inherently subjective and context-dependent.

In paper I and II, we could not directly compare rating scales versus language-based assessment to a direct measure of depression or anxiety. In paper I, we systematically evaluated different language-based assessment response formats and found that both closed-ended (e.g., select words) and open-ended (e.g., text responses) formats demonstrated high convergent validity with traditional rating scales for depression and anxiety. Notably, open-ended formats tended to offer greater convergence with external criteria (e.g., sick leave, health care use), and higher information content. While select words formats achieved high assessment accuracy efficiently, the open-ended text formats provided richer and more personalized data, often equaling or surpassing the external validity of the rating scales. However, a trade-off was noted in terms of test-retest reliability, where rating scales outperformed language-based assessments, possibly due to their constrained format being less sensitive to actual changes in mental state over time.

In paper II, we evaluated the validity of language-based assessments against expert-rated ratings of suicide risk made by expert clinicians. These expert judgments were based on a comprehensive review of participants' self-reported data, including both rating scales and open-ended language responses collected over a 10-week follow-up period. Because the expert assessments themselves incorporated both forms of data, it was not possible to directly compare the assessment performance of language-based assessments and rating scales against an independent ground truth. Instead, the study focused on demonstrating that language-based assessments could approximate these expert-rated assessment ratings by clinical experts with a high degree of accuracy, supporting their criterion validity. The models achieved robust correlations with clinician-rated suicide and self-harm risk, maintained performance in held-out samples, and exhibited clear face validity through predictive language features that aligned with established clinical themes. These findings indicate that language-based assessments can serve not only as valid measurement tools but also as clinically meaningful ways of describing mental health risk, offering exploratory approaches for settings where traditional assessments are less accessible.

Paper III enabled us to compare the ability of language-based assessments versus rating scales to capture changes in affect – demonstrating the superiority of language-based assessments in this setting. We conducted an experimental study to evaluate the causal validity of language-based assessments compared to the Positive and Negative Affect Schedule. Through a mood induction procedure, we found that language-based assessments outperformed the Positive and Negative Affect Schedule scores in classifying participants' emotional states post-induction across all evaluation datasets (cross-validation and holdout samples). Language-based assessments could also describe participants' affective experiences in a highly expressive and individualized manner, as they drew directly on the participants' own words. This allowed for rich contextual interpretations and provided insight into the emotional nuances that traditional rating scales could not capture. For instance, language-based assessments revealed affective language patterns that aligned with each condition's experiential features—such as terms indicating calm or awe in the nature setting, social tension in the mall, or spirituality in the church—underscoring their value in both assessment and interpretation. Thus, language-based assessments were not only more sensitive to experimental manipulations but also very capable of revealing nuanced affective states. This causal validity suggests that language-based assessments may be particularly well-suited for use in intervention-based assessment contexts, such as therapy monitoring (Flemotomos et al., 2022).

Regarding Feasibility and Openness issues, specifically, while language-based assessments require considerably more computational infrastructure and know-how than rating scales, they are scalable and automatable once models are developed and validated. Open-science resources such as the L-BAM Library further reduce technical barriers, allowing non-specialists to apply and interpret language-based assessments with minimal setup. Response formats can also be tailored to suit different time or literacy constraints, as shown in Paper I, which compared response styles ranging from single words to full-text narratives.

Further, language-based assessments rely on language generation, which may disadvantage individuals with limited verbal fluency or impairments in expressive language. Additionally, language responses can introduce privacy concerns due to their potential identifiability and semantic richness. These issues necessitate careful ethical considerations, including anonymization protocols and user-informed consent, as discussed in Chapter 6.

Moreover, while language-based assessments differ from traditional scales in structure, this thesis demonstrates that they can meet key reliability standards—particularly through prospective evaluation and test–retest designs—ensuring that the models are not only valid but also consistent and dependable over time and across samples (see Table 2 also for an overview of reliability concepts). For instance, Paper I applied a test–retest design and reported moderate reliability over a two-week interval, although these correlations were lower than those reported for closed-ended rating scales. This lower stability may initially appear as a limitation, yet it can also reflect a strength: sensitivity to genuine psychological change rather than rigid consistency. Paper III supports this view by demonstrating that language-based assessments captured subtle but systematic changes in affect induced by different environmental conditions more accurately than traditional rating scales. This responsiveness to change suggests that while language-based assessments may show lower test–retest reliability under static conditions, they can provide more sensitive tools for capturing dynamic psychological processes—a quality essential for use in intervention and experimental settings. Together, these findings highlight that reliability, particularly in the context of language-based assessments, must be interpreted in light of the purpose and use case of the assessment.

Potential Implications for Psychological Assessment

The findings of this thesis have several important implications for the science and practice of psychological assessment. First and foremost, the results demonstrate that language-based assessments may offer valid and reliable insights in research settings and potentially complement traditional rating scales. Across multiple studies, language responses assessed core mental health constructs such as depression, anxiety, and suicidality with strong concurrent validity. This suggests that open-ended language, when paired with modern natural language processing, is not only expressive but also measurable in a psychometrically sound way.

Second, the results reveal that language-based assessments offer unique advantages in construct representation. Unlike rating scales, which are constrained by predefined response options, language-based assessments allow individuals to express their psychological states in their own words. This openness results in higher information content, greater sensitivity to subtle emotional shifts, and improved ecological validity. These properties may be particularly valuable in dynamic or personalized assessment contexts, such as ongoing therapy, mood tracking, or digital interventions, where nuanced understanding is critical.

Third, this thesis provides empirical support for the causal validity of language-based assessments. Paper III demonstrated that language-based assessments could detect experimentally induced changes in affect more accurately than closed-ended ratings. This suggests that language-based assessments may offer exploratory utility in tracking psychological change in response to interventions or environmental shifts—an essential feature for monitoring outcomes in clinical and experimental settings.

Finally, by integrating open science principles into the development and dissemination of models, this work supports transparency and collaboration in computational psychology, offering tools for research-oriented development. The L-BAM Library and supporting tools make validated models readily available, lowering the barrier for adoption and encouraging independent validation. This infrastructure helps bridge the gap between technical innovation and applied psychological use, supporting a more cumulative and accountable approach to assessment research.

In this light, language-based assessments do not aim to replace traditional tools or clinical judgment. Instead, they offer a way to enhance how we listen, understand, and respond—at scale, with nuance, and with sensitivity to individual expression.

Their potential strength lies in complementing existing research practices by adding interpretive depth and flexibility to psychological assessment, rather than serving as a competing or standalone alternative.

While language-based assessments offer unique advantages—such as expressiveness, flexibility, and contextual sensitivity—they are best viewed as one component in a broader toolkit. Traditional tools, including rating scales and structured interviews, remain more suitable for certain conditions or populations, especially when responses require focused symptom quantification or when individuals struggle to articulate their experiences in open-ended form. Conversely, language-based assessments may be explored in contexts that benefit from deeper narrative understanding or scalable remote assessments. More research is needed to identify which settings and constructs are best served by language-based methods, and where they might complement existing tools. Validating their use across diverse populations and conditions is essential to ensure fair, effective, and evidence-based use within research settings, while future work should evaluate clinical applicability.

Limitations of the Present Work

While this thesis presents promising findings on the use of AI-driven language-based assessments for evaluating mental health, several limitations must be acknowledged to contextualize the results and guide future research.

Overlap between Depression and Anxiety. A particular challenge for discriminant validity is the close comorbidity and symptom overlap between depression and anxiety. Instruments such as the PHQ-9 and GAD-7 capture related domains, and their high correlations mean that statistical methods like creating difference scores (e.g., subtracting normalized PHQ-9 from GAD-7 scores) are psychometrically unstable, conceptually ambiguous, and difficult to interpret clinically. Although such difference scores have been tested in exploratory contexts, they do not provide strong evidence of construct separation. Instead, careful instrument design offers a more defensible path. For example, the Beck Anxiety Inventory (BAI; Beck, Epstein, Brown, & Steer, 1988) was explicitly developed to emphasize physiological anxiety symptoms and reduce overlap with depressive symptoms, thereby enhancing discriminant validity. Future work on language-based assessments could take inspiration from such design principles: rather than relying on artificial contrasts between highly correlated rating scales, models could be trained to detect theoretically distinctive linguistic markers of depression and anxiety, while acknowledging their frequent co-occurrence in clinical populations.

Sample Generalizability. The empirical studies primarily relied on English or Swedish speaking participants recruited from online platforms panels. Although the samples were diverse in terms of age and psychological symptom severity, they do not fully represent the broader population—particularly with respect to cultural, linguistic, and clinical diversity. This limitation is not merely statistical, but conceptual: it raises questions about whether expressions of distress captured in these models reflect culturally specific norms rather than universal features of mental suffering. For instance, idioms of distress may differ significantly across communities (e.g., somatic expressions in some East Asian cultures [Yap et al., 2024], more I-use in white but not in black [Rai et al., 2024a; Rai et al., 2024b], or spiritual framings in certain religious contexts [Fennig, & Denov, 2025]), and silence or emotional restraint may carry different meanings in collectivist versus individualist societies. As such, models trained on relatively homogenous populations risk misinterpreting, pathologizing, or overlooking expressions of distress that fall outside dominant cultural norms. This limits the generalizability of the models to other languages, cultural groups, and clinical populations not included in the training and evaluation data. Future research should engage with these issues more directly, incorporating cross-cultural data and explicitly testing model performance across diverse subgroups to avoid reinforcing existing biases and disparities in psychological assessment.

Clinical Relevance of the Online Sample in Paper I. While the prospective dataset shows average scores below clinical cut-offs for PHQ-9 and GAD-7, the training sample includes participants with clinically elevated symptoms—specifically, a mean PHQ-9 score of 11.56 and a GAD-7 score of 10.1, both exceeding the standard clinical thresholds (Kroenke et al., 2001; Spitzer et al., 2006). Additionally, participants in the training set report substantially more sick leave, both over the past three months and the past year, compared to the prospective sample. Importantly, although the prospective sample may reflect subclinical levels of distress, language-based depression scores in this group are significantly correlated with functional outcomes such as sick leave. This suggests that the language models may capture clinically relevant patterns even in populations with milder symptoms, thus supporting their broader applicability and external validity. Further research could use cut-offs for screening to increase the clinical relevance of the online sample.

Language and Cultural Specificity. All language data used in model training and evaluation were in Swedish or English, and model performance may differ when applied to other languages or dialects. Since psychological expression is shaped by cultural norms and linguistic conventions, language-based assessment models developed in one language may not transfer well without adaptation and further validation. This highlights the need for multilingual and cross-cultural research to support more inclusive language-based assessments.

Directive Nature of Prompts. A further limitation, particularly relevant to Paper I, concerns the directive nature of the prompts used to elicit responses. Participants were explicitly asked to describe "how depressed they are" and encouraged to use clear and indicative words. While this ensured comparability across response formats, it may also have constrained the openness of the responses. In effect, respondents were instructed to produce language that could be easily mapped onto scale logic (e.g., “use stronger words if you feel more depressed”). This raises the possibility that the apparent richness of the open-ended formats was partly shaped by demand characteristics and self-interpretive framing, rather than by fully spontaneous descriptions of lived experience. Thus, our study may not have captured the full potential of open-ended formats to reveal complexity, ambivalence, or idiosyncrasy in emotional expression. Future research should experiment with alternative prompts—less evaluative and less directive—to examine whether they elicit qualitatively different kinds of language, and whether such responses improve construct validity beyond what can be achieved through scale-like framing.

Scope of Constructs. The thesis focuses on a specific subset of psychological constructs—namely depression, anxiety, and suicide risk. While these are highly relevant and prevalent in both clinical and public health contexts, the broader applicability of language-based assessments to other domains (e.g., psychosis in Dalal et al., 2025; personality disorders in Entwistle, 2023; trauma in Son et al., 2023;) remains unexplored within this work. Similarly, the models were designed for adults, and their validity for children, adolescents, or older adults remains to be established.

Instrumental Use of Contextual Models. I acknowledge that the thesis draws on ecological, interactionist, and dialogical frameworks more as methodological justifications than as fully integrated theoretical anchors. The primary aim was to test whether open-ended language, analyzed with LLMs, can validly assess familiar constructs, which kept the work within existing diagnostic boundaries. This likely reinforced traditional categories rather than reshaping the logic of measurement. Relatedly, language was treated mainly as a quantifiable signal rather than as situated, negotiated expression; a fuller dialogical approach would treat meaning as co-constructed across speakers, turns, and contexts (Hermans, 2001a, 2001b). Accordingly, I should empathize that the contribution is a novel method within an existing paradigm, not a redefinition of constructs or nosology.

Contextual and Situational Variation. Language is inherently context-sensitive (Wynn et al., 2024), and the meaning of words can change depending on situational and interpersonal cues. While the models performed well in structured settings with predefined prompts, their robustness in more naturalistic, unstructured environments (e.g., therapy transcripts, social media posts) remains an open question. Additional testing in clinical settings contexts is necessary to assess whether model assessments retain validity and reliability under varied conditions.

Model Interpretability and Black-Box Risks. Despite efforts to improve interpretability through visualizations and construct-aligned language features, large language models remain complex and often opaque (Liao, & Vaughan, 2024). The decision-making processes of models like RoBERTa (Liu et al., 2019) or GPT (Radford et al., 2018) are not fully transparent, which can hinder clinical adoption and limit trust among users (Navandi et al., under review). This opacity also complicates the detection of model bias or failure cases, especially in high-stakes scenarios such as suicide risk assessment. A supplementary paper (Eijsbroek et al., 2025b), co-authored by the author of this thesis, aims to enhance clarity through various types of word plots, including text projection plots and word clusters derived from topic modeling.

Temporal and Longitudinal Evaluation. Although some test–retest analyses were conducted, more extensive longitudinal evaluations are needed to assess the stability and sensitivity of language-based assessments over time. This includes understanding how language-based scores evolve in response to interventions, life events, or clinical deterioration. Ongoing studies should assess whether these tools are suitable for monitoring change in clinical settings for therapy monitoring.

Vulnerability to Social Desirability Bias. Although language-based assessments offer greater flexibility and nuance compared to traditional rating scales, they are not immune to social desirability bias. Participants may still consciously or unconsciously shape their responses to present themselves in a favorable light (e.g., Salecha et al., 2024), particularly when discussing sensitive topics such as suicidality or depression. This can lead to underreporting of distress or exaggeration of socially acceptable coping strategies (e.g., McDonald, 2008, & Omari et al., 2024), thereby reducing the authenticity of the language data. While language-based assessments allow for more open expression, the desire to conform to perceived social norms remains a challenge that needs to be accounted for in both the design and interpretation of assessments.

Limited by Participants' Insight. The effectiveness of language-based assessments relies heavily on participants' ability to accurately access and describe their own psychological states. While this openness allows for rich, individualized expression, it also assumes a certain level of introspective ability and verbal fluency. In contrast, structured rating scales may assist individuals by offering cognitive scaffolding (Grindheim et al., 2024)—helping them recognize and report experiences they might not identify independently. But more importantly, language-based assessments need not be entirely open or unstructured; they can adopt more focused, symptom-specific prompts (e.g., moving from general mental health to low mood or sleep problems), thereby supporting participants in articulating relevant psychological states more effectively. However, individuals vary widely in their level of self-awareness and insight into their emotional or cognitive experiences (Wilson, & Dunn, 2004). Those with limited insight—such as individuals experiencing alexithymia (i.e., who have difficulty identifying and articulating emotions due to impaired emotional processing and reduced verbal expressiveness; Sikström, et al., 2024), certain personality traits, or early stages of mental illness—

may struggle to provide meaningful or accurate language responses. This limitation is particularly relevant in clinical contexts, where underdeveloped insight may obscure key symptoms or distort self-descriptions, thereby affecting the validity of language-based assessments (Kenny et al., 1994).

Self-Report Anchoring. Although the thesis emphasizes moving beyond traditional rating scales, several validation efforts (particularly in Paper I) still rely on self-report instruments (e.g., Patient Health Questionnaire-9) as proxies for ground truth. These anchors themselves are limited by introspective accuracy. While Paper II introduces expert-rated ratings and Paper III uses experimental manipulation for causal testing, future work should further prioritize external, behavioral, or longitudinal benchmarks.

No (yet) Norms and Cutoffs. A notable advantage of traditional rating scales is the availability of well-established norms and clinically validated cutoffs, which aid interpretation and decision-making in both research and applied settings. In contrast, language-based assessments currently lack standardized reference values. This absence makes it difficult to determine what constitutes "mild," "moderate," or "severe" symptomatology from language alone, potentially limiting clinical utility. Without normative distributions or established thresholds, practitioners and researchers may face uncertainty when interpreting language-based assessments outputs—particularly in high-stakes contexts such as diagnosis, treatment planning, or risk assessment. Developing robust norms and validated cutoffs, ideally tailored to different populations and use cases, remains a key step toward integrating language-based assessments into routine practice.

Lack of Information, Including Context, Voice. One notable limitation of current language-based assessments is their reliance solely on written text, without access to contextual cues such as tone of voice (e.g., Pell et al., 2009), facial expressions (e.g., Niedenthal et al., 2002), or interactional dynamics. Contextual nuances—such as irony, negation, or emotional intensity—can critically shape the intended meaning of a response and may be misinterpreted when text is analyzed in isolation. Additionally, responses are typically evaluated without information about the social or environmental context in which they were generated. Future research should explore integrating other signals—such as prosody, or surrounding discourse—to enhance the sensitivity and accuracy of language-based models in real-world applications.

Together, these limitations highlight the importance of interpreting language not as a standalone indicator but as a product of dynamic person–context interactions. The interactionist model (Judge, & Zapata, 2015; Tett, & Burnett, 2003) provides a valuable lens here, emphasizing that language use reflects the interplay between internal states, social norms, and situational factors. This perspective reinforces the need for caution when generalizing results from language-based assessments and calls for future models to account for these contextual influences—either through multimodal data, adaptive prompts, or situational framing—in order to produce more accurate, equitable, and ecologically valid assessments.

Future Directions

The findings of this thesis point to a range of promising future directions for advancing language-based assessments in psychological science and practice. As the field evolves, several areas of methodological, theoretical, and applied development warrant attention.

Ensuring Context-Specific Validation Before Clinical Deployment. While the current thesis contributes to the growing body of evidence that language-based assessments, particularly those powered by large language models, can yield strong validity across a variety of metrics—including convergent, discriminant, and external validity—important limitations remain regarding their readiness for widespread clinical or population-level deployment. Specifically, although the included studies show robust psychometric performance in research settings, there is not yet a validated "one-size-fits-all" language model that can be universally applied across diverse populations, languages, or clinical conditions.

The current work should therefore be understood as contributing support for large language-model-based assessments as a class of techniques rather than as endorsing any specific model instance for general use (see Kjell et al., 2024). Each individual model must undergo thorough, context-specific evaluation before being used for clinical or diagnostic purposes—just as is required for traditional rating scales. This includes testing for reliability, construct validity, and potential biases within the specific population and setting in which the tool is intended to be used.

To date, relatively few language-based assessment models have been validated in clinical deployment scenarios such as therapy, screening, or diagnostics, and most evaluations remain confined to cross-sectional or survey-based designs. As emphasized in recent reviews (e.g., Eichstaedt et al., 2018; Soni et al., 2022), future work must expand the evidence base through longitudinal, ecologically valid studies that assess real-world applicability. Moreover, dynamic validation is necessary—such as evaluating whether language models can detect changes over time or predict symptom trajectories.

In sum, while this thesis provides strong support for the potential of large language-model-based assessments, it does not assume that all implementations of these models are inherently valid or trustworthy. Responsible deployment will require targeted validation, careful monitoring, and ongoing refinement in alignment with evolving ethical and regulatory standards.

Broader Construct Coverage and Clinical Utility. Future work should expand beyond depression, anxiety, and suicide risk to encompass a broader spectrum of psychological constructs (Vu et al., 2024). This includes conditions where individuals' subjective experience plays a central role in diagnosis and treatment—such as depression, anxiety, and suicidality—where self-report remains a primary source of insight. In contrast, other conditions like attention-related disorders or neurocognitive impairments may benefit more from behavioral or task-based assessments (Flemotomos et al., 2022). Extending language-based assessments into these areas may involve developing new prompt strategies and embedding frameworks tailored to specific symptom profiles. Moreover, clinical trials should test the utility of language-based assessments as adjuncts to diagnosis, treatment planning, and therapy monitoring in real-world mental health services, preferably using randomized controlled trials (RCTs).

Beyond Epistemological Circularity. To move beyond the limitation of epistemological circularity, future research should place stronger emphasis on external validation. One path is to test whether language-based assessments converge with best-estimate ratings derived from longitudinal expert reviews that integrate multiple data sources, as in the longitudinal expert data (LED) approach used in Paper II. Another is to evaluate predictive value with respect to behavioral and clinical outcomes such as treatment response, functional impairment, or relapse and remission trajectories. By extending validation beyond self-report scales, language-based methods can be assessed not merely for their ability to reproduce existing measures, but for their potential to add unique explanatory and predictive value to psychological assessment.

Cross-Linguistic and Cross-Cultural Validation. A critical next step is the development and evaluation of language-based assessments in multiple languages and cultural contexts (Liu et al., 2024). This involves both technical adaptations (e.g., training models on multilingual corpora) and cultural considerations (e.g., understanding how psychological states are expressed differently across linguistic traditions). Open collaboration across research teams globally will be vital for building a library of culturally responsive language-based assessment tools that are generalizable and equitable.

Integration with Longitudinal and Behavioral Data. To better assess the stability and predictive utility of language-based assessments, future studies should incorporate longitudinal designs that track participants over time (Burkhardt et al., 2021). This would allow researchers to examine how changes in language use relate to clinical outcomes, behavioral events (e.g., hospitalization), or life changes. Integrating language-based assessments with passive digital data (e.g., smartphone use, physiological monitoring) may also support the development of multimodal digital phenotyping approaches for mental health.

Real-Time and Adaptive Assessment. Advances in computing power and mobile technology open opportunities for language-based assessments to be used in real-time, ecological momentary assessments (EMA, in Yin et al., 2024). Adaptive systems could respond to a user's input by tailoring follow-up prompts or suggesting interventions based on current language use. These systems could enhance mental health monitoring in outpatient care, support just-in-time interventions, or provide scalable tools for early detection of risk.

Ethical Governance and Clinical Guidelines. As language-based assessments move closer to implementation, future work must engage more deeply with the ethical, legal, and professional guidelines needed for safe use (Lekadir et al., 2025). This includes establishing standards for model documentation, fairness auditing, user consent, and clinician involvement. Partnerships with regulatory bodies, ethics boards, and professional associations will be crucial for integrating language-based assessments into evidence-based frameworks and clinical decision-making.

Toward Non-Instrumental Use. Future work could move beyond predictive replication of existing tools by (a) modeling dialogical processes and discourse features in patient–clinician (and peer) interactions (e.g., turn-taking, repair, hedging, silence), (b) embedding multi-level context (speaker, relationship, setting, and broader social conditions) consistent with ecological systems theory (Bronfenbrenner, 1979), and (c) testing language models against dimensional and mechanism-based alternatives to categorical diagnosis (Cuthbert, 2022; Insel & Cuthbert, 2010; HiTOP in Kotov et al., 2017). The goal is to let open-ended language surface emergent constructs and relational/process markers that complement (rather than replace) traditional categories.

Theoretical Refinement. Finally, future research should continue to refine the theoretical underpinnings of language-based assessments (Boyd, & Schwartz, 2001). While current work draws on constructs like attention, emotion, and linguistic behavior, there is room to more clearly define what is being measured by language—and how. Bridging psycholinguistics, social psychology, and computational modeling may yield better conceptual clarity, especially in understanding the interplay between context, expression, and mental state.

In sum, language-based assessments represent a rapidly advancing frontier in psychological science. With rigorous validation, ethical foresight, and collaborative development, these tools have the potential to significantly improve how we understand, measure, and respond to mental health—both in research and in practice.

Concluding Remarks

This thesis set out to explore whether and how natural language—an inherently human mode of expression—can be harnessed through artificial intelligence to assess mental health in ways that are valid, reliable, and ethically sound. Across four empirical studies, the work has demonstrated that language-based assessments are not only technically feasible but also scientifically robust. By transforming open-ended responses into interpretable psychological insights, language-based assessments provide a flexible and scalable complement to traditional assessment methods.

The findings show that language-based assessments can match or exceed the performance of conventional rating scales across multiple domains, including depression, anxiety, and suicidality risk. They are sensitive to momentary psychological changes, align well with clinician assessments, and offer greater descriptive richness—especially in open-ended formats. These results underscore the potential of language as both a source of data and a vehicle for capturing the nuanced realities of psychological experience.

However, the promise of language-based assessments is not just technological. It reflects a broader shift in how we think about psychological measurement: from fixed formats to expressive language, from static questionnaires to dynamic models, and from isolated assessments to integrated, responsive systems. This transformation invites us to reimagine assessment as a conversation rather than a checklist—as an interaction that values context, voice, and individual variability.

Looking ahead, the successful integration of language-based assessments into psychological science will depend on more than continued innovation. It will require interdisciplinary collaboration, thoughtful validation, and deep ethical engagement. This thesis contributes foundational steps in that direction—offering methods, models, and frameworks that others can hopefully build upon.

References

Adams, R., Haroz, E. E., Rebman, P., Suttle, R., Grosvenor, L., Bajaj, M., Dayal, R. R., Maggio, D., Kettering, C. L., & Goklish, N. (2024). Developing a suicide risk model for use in the Indian Health Service. *Npj Mental Health Research*, *3*(1), 47.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing.* Washington, DC: AERA.

Armstrong, N., & Byrom, N. (2025). An anthropological critique of psychiatric rating scales. *BJPsych Advances*, *31*(2), 73–81.

Arroll, B., Goodyear-Smith, F., Crengle, S., Gunn, J., Kerse, N., Fishman, T., Falloon, K., & Hatcher, S. (2010). Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. *The Annals of Family Medicine*, *8*(4), 348–353.

American Psychiatric Association. (2022). *Diagnostic and Statistical Manual of Mental Disorders: Fifth Edition, Text Revision (DSM-5-TR)*. American Psychiatric Association Publishing.

Bakker, D., & Rickard, N. (2018). Engagement in mobile phone app for self-monitoring of emotional wellbeing predicts changes in mental health: MoodPrism. *Journal of affective disorders*, *227*, 432-442.

Barrett, L. F., Lindquist, K. A., & Gendron, M. (2007). Language as context for the perception of emotion. *Trends in cognitive sciences*, *11*(8), 327-332.

Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: psychometric properties. *Journal of consulting and clinical psychology*, *56*(6), 893.

Bender, T. W., Fitzpatrick, S., Hartmann, M. A., Hames, J., Bodell, L., Selby, E. A., & Joiner, T. E. (2019). Does it hurt to ask? An analysis of iatrogenic risk during suicide risk assessment. *Neurology, Psychiatry and Brain Research*, *33*, 73–81.

Berry-Blunt, A. K., Holtzman, N. S., Donnellan, M. B., & Mehl, M. R. (2021). The story of “I” tracking: Psychological implications of self-referential language use. *Social and Personality Psychology Compass*, *15*(12), e12647.

Bhatia, S., & Aka, A. (2022). Cognitive modeling with representations from large-scale digital data. *Current Directions in Psychological Science*, *31*(3), 207–214.

Bihu, R. (2020). Using unstructured interviews in educational and social science research: The process, opportunity and difficulty. *Global Scientific Journals, GSJ*, *8*(10).

Bjureberg, J., Dahlin, M., Carlborg, A., Edberg, H., Haglund, A., & Runeson, B. (2022). Columbia-Suicide Severity Rating Scale Screen Version: Initial screening for suicide risk in a psychiatric emergency department. *Psychological Medicine*, *52*(16), 3904–3912.

Black, D. W., Coryell, W. H., Crowe, R. R., McCormick, B., Shaw, M. C., & Allen, J. (2014). A direct, controlled, blind family study of DSM-IV pathological gambling. *J Clin Psychiatry*, *75*(3), 215–221.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, *111*(4), 1061–1071.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., Lijmer, J. G., Moher, D., Rennie, D., & others. (2003). *Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative.*

Boyd, R. L., & Markowitz, D. M. (2025). Verbal behavior and the future of social science.*American Psychologist, 80*(3), 411–433.

Boyd, R. L., & Pennebaker, J. W. (2015). A way with words: Using language for psychological science in the modern era. In *Consumer psychology in a social media world* (pp. 222–236). Routledge.

Boyd, R. L., & Pennebaker, J. W. (2017). Language-based personality: A new approach to personality in a digital world. *Current Opinion in Behavioral Sciences*, *18*, 63–68.

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.

Brandt, P. A. (2022). Saussure's Prolegomena—Toward a Semiotics of the Mind. *Language and Semiotic Studies*, *8*(1), 91–104.

Bredström, A. (2019). Culture and context in mental health diagnosing: Scrutinizing the DSM-5 revision. *Journal of Medical Humanities*, *40*(3), 347–363.

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Harvard University Press.

Brunswik, E. (1949). DISCUSSION: REMARKS ON FUNCTIONALISM IN PERCEPTION. *Journal of Personality*, *18*(1), 56–65.

Brunswik, E. (2023). *Perception and the Representative Design of Psychological Experiments*. University of California Press.

Burkhardt, H. A., Alexopoulos, G. S., Pullmann, M. D., Hull, T. D., Areán, P. A., & Cohen, T. (2021). Behavioral activation and depression symptomatology: Longitudinal assessment of linguistic indicators in text-based therapy sessions. *Journal of Medical Internet Research*, *23*(7), e28244.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81.

Campbell, D. T., & Stanley, J. C. (2015). *Experimental and quasi-experimental designs for research*. Ravenio books.

Cardamone, N. C., Olfson, M., Schmutte, T., Ungar, L., Liu, T., Cullen, S. W., Williams, N. J., & Marcus, S. C. (2025). Classifying Unstructured Text in Electronic Health Records for Mental Health Prediction Models: Large Language Model Evaluation Study. *JMIR Medical Informatics*, *13*(1), e65454.

Cerda, P., Varoquaux, G., & Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. *Machine Learning*, *107*(8), 1477–1494.

Cohen, J., Wright-Berryman, J., Rohlfs, L., Wright, D., Campbell, M., Gingrich, D., Santel, D., & Pestian, J. (2020). A feasibility study using a machine learning suicide risk prediction model based on open-ended interview language in adolescent therapy sessions. *International Journal of Environmental Research and Public Health*, *17*(21), 8187.

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Cooper, C. (2023). *An introduction to psychometrics and psychological assessment: Using, interpreting and developing tests*. Routledge.

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 51–60.

Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. (2015). From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 1–10.

Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights*, *10*, 1178222618792860.

Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., ... & de Oliveira, N. (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns*, *4*(10).

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

Cukrowicz, K. C., Smith, P. N., & Poindexter, E. K. (2010). The Effect of Participating in Suicide Research: Does Participating in a Research Protocol on Suicide and Psychiatric Symptoms Increase Suicide Ideation and Attempts? *Suicide and Life-Threatening Behavior*, *40*(6), 535–543.

Cuthbert, B. N. (2022). The RDoC framework: Facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. *Psychophysiology, 59*(4), e14053.

Dalal, T. C., Liang, L., Silva, A. M., Mackinley, M., Voppel, A., & Palaniyappan, L. (2025). Speech based natural language profile before, during and after the onset of psychosis: A cluster analysis. *Acta Psychiatrica Scandinavica*, *151*(3), 332–347.

Davis, M. A. C., Spriggs, A., Rodgers, A., & Campbell, J. (2018). The effects of a peer-delivered social skills intervention for adults with comorbid down syndrome and autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *48*, 1869–1885.

Dazzi, T., Gribble, R., Wessely, S., & Fear, N. T. (2014). Does asking about suicide and related behaviours induce suicidal ideation? What is the evidence?. *Psychological medicine*, *44*(16), 3361-3363.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), 128–137.

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016). Discovering shifts to suicidal ideation from mental health content in social media. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2098–2110.

DeCou, C. R., & Schumann, M. E. (2018). On the iatrogenic risk of assessing suicidality: A meta-analysis. *Suicide and Life-Threatening Behavior*, *48*(5), 531–543.

De Manuel, A., Delgado, J., Parra Jounou, I., Ausín, T., Casacuberta, D., Cruz, M., Guersenzvaig, A., Moyano, C., Rodríguez-Arias, D., Rueda, J., & others. (2023). Ethical assessments and mitigation strategies for biases in AI-systems used during the COVID-19 pandemic. *Big Data & Society*, *10*(1), 20539517231179199.

DeJonckheere, M., & Vaughn, L. M. (2019). Semistructured interviewing in primary care research: A balance of relationship and rigour. *Family Medicine and Community Health*, *7*(2), e000057.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., & others. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), 688–701.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

DeYoung, C. G., Chmielewski, M., Clark, L. A., Condon, D. M., Kotov, R., Krueger, R. F., Lynam, D. R., Markon, K. E., Miller, J. D., Mullins-Sweatt, S. N., & others. (2022). The distinction between symptoms and traits in the Hierarchical Taxonomy of Psychopathology (HiTOP). *Journal of Personality*, *90*(1), 20–33.

Dudda, L., Kormann, E., Kozula, M., DeVito, N. J., Klebel, T., Dewi, A. P., Spijker, R., Stegeman, I., Van den Eynden, V., Ross-Hellauer, T., & others. (2025). Open science interventions to improve reproducibility and replicability of research: A scoping review. *Royal Society Open Science*, *12*(4), 242057.

Dumas, D., Greiff, S., & Wetzel, E. (2025). TEN GUIDELINES FOR SCORING PSYCHOLOGICAL ASSESSMENTS USING ARTIFICIAL INTELLIGENCE. *European Journal of Psychological Assessment*.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., Jha, S., Agrawal, M., Dziurzynski, L. A., Sap, M., & others. (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., Asch, D. A., & 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. (2025a). The LEADING guideline: Reporting standards for expert panel, best-estimate diagnosis, and longitudinal expert all data (LEAD) methods. *Comprehensive Psychiatry*, 152603.

Eijsbroek, V. C., Nilsson, A., Gu, Z., Wiebel, C., Ganesan, A. V., Kjell, K., … Kjell, O. N. E. (2025b). Multiple Methods for Visualizing Human Language: A Tutorial for Social and Behavioural Scientists. https://doi.org/10.31234/osf.io/nxfvr_v1

Ekstrand Odd, K. (2024). Assessing meaningful change in mental health using large language models: An anchor-based approach (Master's thesis). [Lund University].

Entwistle, C. (2023). *Personality Dysfunction Manifest in Words: Understanding Personality Pathology Using Computational Language Analysis* [PhD Thesis]. Lancaster University (United Kingdom).

Fennig, M., & Denov, M. (2025). Exploring "language of suffering": Idioms of distress among Eritrean refugees living in Israel. *Qualitative Health Research, 35*(4-5), 476-490.

Frank, J. D., Frank, J. B., & Wampold, B. E. (2025). *Persuasion and healing: A comparative study of psychotherapy*. JhU Press.

Feder, K. M., Rahr, H. B., Lautrup, M. D., Egebæk, H. K., Christensen, R., & Ingwersen, K. G. (2022). Effectiveness of an expert assessment and individualised treatment compared with a minimal home-based exercise program in women with late-term shoulder impairments after primary breast cancer surgery: Study protocol for a randomised controlled trial. *Trials*, *23*(1), 701.

Feuerriegel, S., Maarouf, A., Bär, D., Geissler, D., Schweisthal, J., Pröllochs, N., Robertson, C. E., Rathje, S., Hartmann, J., Mohammad, S. M., & others. (2025). Using natural language processing to analyse text data in behavioural science. *Nature Reviews Psychology*, 1–16.

First, M. B. (2014). Structured clinical interview for the DSM (SCID). *The Encyclopedia of Clinical Psychology*, 1–6.

Fitzner, K. (2007). Reliability and validity a quick review. *The Diabetes Educator*, *33*(5), 775–780.

Flemotomos, N., Martinez, V. R., Chen, Z., Singla, K., Ardulov, V., Peri, R., Caperton, D. D., Gibson, J., Tanana, M. J., Georgiou, P., & others. (2022). Automated evaluation of psychotherapy skills using speech and language technologies. *Behavior Research Methods*, *54*(2), 690–711.

Flusberg, S. J., Holmes, K. J., Thibodeau, P. H., Nabi, R. L., & Matlock, T. (2024). The psychology of framing: How everyday language shapes the way we think, feel, and act. *Psychological Science in the Public Interest*, *25*(3), 105–161.

Galton, F. (1884). Measurement of Character. *Fortnightly Review*, *36*, 179–185.

Gardner, F. (2000). Methodological issues in the direct observation of parent–child interaction: Do observational findings reflect the natural behavior of participants? *Clinical Child and Family Psychology Review*, *3*, 185–198.

Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T. J., Dobson, R. J., & Dutta, R. (2017). Characterisation of mental health conditions in social media using Informed Deep Learning. *Scientific reports*, *7*(1), 1-11

Goanta, C., Aletras, N., Chalkidis, I., Ranchordas, S., & Spanakis, G. (2023). Regulation and NLP (RegNLP): Taming Large Language Models. *arXiv Preprint arXiv:2310.05553*.

Gould, M. S., Marrocco, F. A., Kleinman, M., Thomas, J. G., Mostkoff, K., Cote, J., & Davies, M. (2005). Evaluating iatrogenic risk of youth suicide screening programs: A randomized controlled trial. *JAMA*, *293*(13), 1635–1643.

Grahek, I., Schaller, M., & Tackett, J. L. (2021). Anatomy of a psychological theory: Integrating construct-validation and computational-modeling methods to advance theorizing. *Perspectives on Psychological Science*, *16*(4), 803–815.

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

Grindheim, Ø., McAleavey, A., Iversen, V., Moltu, C., Tømmervik, K., Govasmark, H., & Brattland, H. (2024). Response processes for patients providing quantitative self-report data: A qualitative study. *Quality of Life Research*, 1–13.

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

Gu, Z., Kjell, K., Schwartz, H. A., & Kjell, O. (2025). Natural language response formats for assessing depression and worry with large language models: A sequential evaluation with model pre-registration. *Assessment*. Advance online publication.

Guevara, M., Chen, S., Thomas, S., Chaunzwa, T. L., Franco, I., Kann, B. H., Moningi, S., Qian, J. M., Goldstein, M., Harper, S., & others. (2024). Large language models to identify social determinants of health in electronic health records. *NPJ Digital Medicine*, *7*(1), 6.

Haibe-Kains, B., Adam, G. A., Hosny, A., Khodakarami, F., Cesare 42, M. A. Q. C. (MAQC) S. B. of D. S. T. 35 K. R. 36 S. S.-A. 37 T. W. 35 W. R. D. 38 M. C. E. 39 J. W. 40 D. J. 41 F., Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., & others. (2020). Transparency and reproducibility in artificial intelligence. *Nature*, *586*(7829), E14–E16.

Hales, A. H., Wesselmann, E. D., & Hilgard, J. (2019). Improving psychological science through transparency and openness: An overview. *Perspectives on Behavior Science*, *42*, 13–31.

Harrer, S. (2023). Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*, *90*.

He, T., Fu, G., Yu, Y., Wang, F., Li, J., Zhao, Q., Song, C., Qi, H., Luo, D., Zou, H., & others. (2023). Towards a psychological generalist ai: A survey of current applications of large language models and future prospects. *arXiv Preprint arXiv:2312.04578*.

Henson, R. K. (2001). Understanding Internal Consistency Reliability Estimates: A Conceptual Primer on Coefficient Alpha. *Measurement and Evaluation in Counseling and Development*, *34*(3), 177–189.

Hermans, H. J. (2001a). The dialogical self: Toward a theory of personal and cultural positioning. *Culture & psychology*, *7*(3), 243-281.

Hermans, H. J. (2001b). The construction of a personal position repertoire: Method and practice. *Culture & Psychology*, *7*(3), 323-366.

Hom, M. A., Stanley, I. H., Rogers, M. L., Gallyer, A. J., Dougherty, S. P., Davis, L., & Joiner, T. E. (2018). Investigating the iatrogenic effects of repeated suicidal ideation screening on suicidal and depression symptoms: A staggered sequential study. *Journal of Affective Disorders, 232*, 139–142.

Huck, S. W. (2012). Reading statistics and research. *Boston, USA*, *4*, 103–112.

Hull, J. G., Van Treuren, R. R., Ashford, S. J., Propsom, P., & Andrus, B. W. (1988). Self-consciousness and the processing of self-relevant information. *Journal of Personality and Social Psychology*, *54*(3), 452.

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

Jackson, J. C., Watts, J., List, J.-M., Puryear, C., Drabble, R., & Lindquist, K. A. (2022). From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science*, *17*(3), 805–826.

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2021). Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv Preprint arXiv:2110.15621*.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.

Judge, T. A., & Zapata, C. P. (2015). The person–situation debate revisited: Effect of situation strength and trait activation on the validity of the Big Five personality traits in predicting job performance. *Academy of Management Journal*, *58*(4), 1149–1179.

Katchapakirin, K., Wongpatikaseree, K., Yomaboot, P., & Kaewpitakkun, Y. (2018). Facebook social media for depression detection in the Thai community. *2018 15th International Joint Conference on Computer Science and Software Engineering (Jcsse)*, 1–6.

Kaźmierczak, I., Jakubowska, A., Pietraszkiewicz, A., Zajenkowska, A., Lacko, D., Wawer, A., & Sarzyńska-Wawer, J. (2024). Natural language sentiment as an indicator of depression and anxiety symptoms: A longitudinal mixed methods study1. *Cognition and Emotion*, 1–10.

Kenny, D. A., Albright, L., Malloy, T. E., & Kashy, D. A. (1994). Consensus in interpersonal perception: Acquaintance and the big five. *Psychological Bulletin*, *116*(2), 245.

Kessler, R. C., & Üstün, T. B. (2004). The world mental health (WMH) survey initiative version of the world health organization (WHO) composite international diagnostic interview (CIDI). *International Journal of Methods in Psychiatric Research*, *13*(2), 93–121.

Kivelä, L. M. M., Fiß, F., Van der Does, W., & Antypa, N. (2024). Examination of acceptability, feasibility, and iatrogenic effects of ecological momentary assessment (EMA) of suicidal ideation. *Assessment*, *31*(6), 1292-1308.

Kjell, O., 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., Ganesan, A. V., Boyd, R., Oltmanns, J. R., Rivero, A., Feltman, S., Carr, M. A., Luft, B., Kotov, R., & Schwartz, H. A. (*in progress*). *Demonstrating High Validity of a New AI-Language Assessment of PTSD: A Sequential Evaluation with Model Pre-registration*.

Kjell, O., Giorgi, S., & Schwartz, H. A. (2023a). The text-package: An R-package for analyzing and visualizing human language using natural language processing and transformers. *Psychological Methods*, *28*(6), 1478.

Kjell, O., Giorgi, S., Schwartz, H. A., & Eichstaedt, J. C. (2023b). Towards well-being measurement with social media across space, time and cultures: Three generations of progress. *World Happiness Report*, 131–162.

Kjell, O. N. E., Kjell, K., & Schwartz, H. A. (2024). Beyond rating scales: With targeted evaluation, large language models are poised for psychological assessment. *Psychiatry Research*, *333*, 115667.

Kjell, O. N., 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., 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.

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., ... & Zimmerman, M. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of abnormal psychology*, *126*(4), 454.

Krieger, V., Magallón Neri, E. M., & Amador, J. A. (2024). *Ethics and artificial intelligence in psychological assessment.*

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613.

Kroenke, K., Stump, T. E., Chen, C. X., Kean, J., Damush, T. M., Bair, M. J., ... & Monahan, P. O. (2021). Responsiveness of PROMIS and Patient Health Questionnaire (PHQ) Depression Scales in three clinical trials. *Health and Quality of Life Outcomes*, *19*(1), 41.

Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. *arXiv Preprint arXiv:1906.07337*.

Insel, T. R., & Cuthbert, B. N. (2009). Endophenotypes: bridging genomic complexity and disorder heterogeneity. *Biological psychiatry, 66*(11), 988-989.

Lawrence, H. R., Schneider, R. A., Rubin, S. B., Matarić, M. J., McDuff, D. J., & Bell, M. J. (2024). The opportunities and risks of large language models in mental health. *JMIR Mental Health*, *11*(1), e59479.

Leidner, J. L., & Plachouras, V. (2017). Ethical by design: Ethics best practices for natural language processing. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 30–40.

Lekadir, K., Frangi, A. F., Porras, A. R., Glocker, B., Cintas, C., Langlotz, C. P., Weicken, E., Asselbergs, F. W., Prior, F., Collins, G. S., & others. (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *Bmj*, *388*.

Levitt, H. M. (2001). Sounds of silence in psychotherapy: The categorization of clients' pauses. *Psychotherapy Research*, *11*(3), 295-309.

Levitt, H. M., Bamberg, M., Creswell, J. W., Frost, D. M., Josselson, R., & Suárez-Orozco, C. (2018). Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 26.

Liao, Q. V., & Vaughan, J. W. (2024). AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review*, *Special Issue 5*.

Lindquist, K. A., & Gendron, M. (2013). What's in a word? Language constructs emotion perception. *Emotion Review*, *5*(1), 66–71.

Linton, M.-J. A., Jelbert, S., Kidger, J., Morris, R., Biddle, L., & Hood, B. (2021). Investigating the use of electronic well-being diaries completed within a psychoeducation program for university students: Longitudinal text analysis study. *Journal of Medical Internet Research*, *23*(4), e25279.

Lison, P., Pilán, I., Sánchez, D., Batet, M., & Øvrelid, L. (2021). Anonymisation models for text data: State of the art, challenges and future directions. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4188–4203.

Liu, C. C., Gurevych, I., & Korhonen, A. (2024). Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *arXiv Preprint arXiv:2406.03930*.

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

Longitudinal Construct Validity: Establishment of Clinical Meaning in Patient Evaluative Instruments. (n.d.). *38*(9).

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, *48*(2), 233-245.

Low, D. M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., & Ghosh, S. S. (2020). Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of Medical Internet Research*, *22*(10), e22635.

McAdams, D. P. (2001). The psychology of life stories. *Review of general psychology*, *5*(2), 100-122.

Malau-Aduli, B. S., Walls, J., & Zimitat, C. (2012). Validity, reliability and equivalence of parallel examinations in a university setting. *Creative Education*, *3*(6), 923–930.

Mangalik, S., Eichstaedt, J. C., Giorgi, S., Mun, J., Ahmed, F., Gill, G., V. Ganesan, A., Subrahmanya, S., Soni, N., Clouston, S. A., & others. (2024). Robust language-based mental health assessments in time and space through social media. *NPJ Digital Medicine*, *7*(1), 109.

Marcusson-Clavertz, D., Kjell, O. N., Persson, S. D., & Cardeña, E. (2019). Online validation of combined mood induction procedures. *PloS One*, *14*(6), e0217848.

Margolin, G., Oliver, P. H., Gordis, E. B., O'hearn, H. G., Medina, A. M., Ghosh, C. M., & Morland, L. (1998). The nuts and bolts of behavioral observation of marital and family interaction. *Clinical Child and Family Psychology Review*, *1*, 195–213.

Matheson, G. J. (2019). We need to talk about reliability: Making better use of test-retest studies for study design and interpretation. *PeerJ*, *7*, e6918.

Mathias, C. W., Furr, R. M., Sheftall, A. H., Hill-Kapturczak, N., Crum, P., & Dougherty, D. M. (2012). What's the Harm in Asking About Suicidal Ideation? *Suicide and Life-Threatening Behavior*, *42*(3), 341–351.

McDonald, J. D. (2008). Measuring personality constructs: The advantages and disadvantages of self-reports, informant reports and behavioural assessments. *Enquire*, *1*(1), 1–19.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 276–282.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence* (Vol. 1, No. 1). Minneapolis: University of Minnesota Press.

Menold, N., Wolf, C., & Bogner, K. (2018). Design aspects of rating scales in questionnaires. In *Mathematical Population Studies* (Vol. 25, Issue 2, pp. 63–65). Taylor & Francis.

Meskó, B., & Topol, E. J. (2023). The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digital Medicine*, *6*(1), 120.

Mesquiti, S., Cosme, D., Falk, E., Nook, E., & Burns, S. M. (2025). Predicting Psychological and Subjective Well-being through Language-based Assessment. *OSF Preprints.* https://osf.io/rfq8p_v1

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, *56*(2), 128.

Miller, D. J., Spengler, E. S., & Spengler, P. M. (2015). A meta-analysis of confidence and judgment accuracy in clinical decision making. *Journal of Counseling Psychology*, *62*(4), 553.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679.

Mohajan, H. K. (2017). Two criteria for good measurements in research: Validity and reliability. *Annals of Spiru Haret University. Economic Series*, *17*(4), 59–82.

Mokkink, L., Terwee, C., & de Vet, H. (2021). Key concepts in clinical epidemiology: responsiveness, the longitudinal aspect of validity. *Journal of clinical epidemiology*, *140*, 159-162.

Mokkink, L. B., De Vet, H., Diemeer, S., & Eekhout, I. (2023). Sample size recommendations for studies on reliability and measurement error: An online application based on simulation studies. *Health Services and Outcomes Research Methodology*, *23*(3), 241–265.

Monson, E., Lonergan, M., Caron, J., & Brunet, A. (2016). Assessing trauma and posttraumatic stress disorder: Single, open-ended question versus list-based inventory. *Psychological Assessment*, *28*(8), 1001.

Mueller, A. E., & Segal, D. L. (2015). Structured versus semistructured versus unstructured interviews. *The Encyclopedia of Clinical Psychology*, 1–7.

Mühlhoff, R. (2023). Predictive privacy: Collective data protection in the context of artificial intelligence and big data. *Big Data & Society*, *10*(1), 20539517231166886.

Navandi, D., Kjell, O. N. E., Kjell, K., Eijsbroek, V., Wiebel, C., Stade, E., & Molendijk, M. (*under review*). Mental Health Assessment Methods and Attitudes in Clinical Practices. *OSF Preprints*. https://osf.io/w7mtz

Niedenthal, P. M., Brauer, M., Robin, L., & Innes-Ker, Å. H. (2002). Adult attachment and the perception of facial expression of emotion. *Journal of Personality and Social Psychology*, *82*(3), 419.

Odermatt, S. D., Grieder, S., Schweizer, F., Bünger, A., & Grob, A. (2025). The Role of Language Aspects in the Assessment of Cognitive and Developmental Functions in Children: An Analysis of the Intelligence and Development Scales–2. *Assessment*, 10731911251315027.

Omari, A., Mahadevan, R., & Elfitasari, D. (2024). Social Desirability and Its Effect on Help-Seeking: Mediated by Shame. *Journal of Applied Youth Psychological Studies*, *10*(2), 45–60.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Palan, S., & Schitter, C. (2018). Prolific. Ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Pallant, J. (2010*). SPSS Survival Manual: A Step by Step Guide to Data Analysis Using IBM SPSS*. Australia: Routledge.

Pedersen, G., Karterud, S., Hummelen, B., & Wilberg, T. (2013). The impact of extended longitudinal observation on the assessment of personality disorders. *Personality and Mental Health*, *7*(4), 277–287.

Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. A. (2009). Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior*, *33*, 107–120.

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological Science*, *8*(3), 162–166.

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.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001), 2001.

Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME Standards for Educational and Psychological Testing? *Educational Measurement: Issues and Practice*, *33*(4), 4–12.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *International Conference on Machine Learning*, 28492–28518.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., & others. (2018). *Improving language understanding by generative pre-training*. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

Rai, S., Stade, E. C., Giorgi, S., Francisco, A., Ungar, L. H., Curtis, B., & Guntuku, S. C. (2024a). Key language markers of depression on social media depend on race. *Proceedings of the National Academy of Sciences, 121*(14), e2319837121.

Rai, S., Stade, E. C., Giorgi, S., Francisco, A., Ganesan, A. V., Ungar, L. H., ... & Guntuku, S. C. (2024b). Reply to Wang: Clarifying model performance and language markers of depression across races. Proceedings of the National Academy of *Sciences, 121*(31), e2410449121.

Raju, N. S., & Guttman, I. (1965). A New Working Formula for the Split-Half Reliability Model. *Educational and Psychological Measurement*, *25*(4), 963–967.

Rao, R., Ganesan, A., Kjell, O., Luby, J., Raghavan, A., Feltman, S., ... & Schwartz, H. A. (2025). WhiSPA: Semantically and Psychologically Aligned Whisper with Self-Supervised Contrastive and Student-Teacher Learning. *arXiv preprint arXiv:2501.16344*.

Reas, D. L., Rø, Ø., Karterud, S., Hummelen, B., & Pedersen, G. (2013). Eating disorders in a large clinical sample of men and women with personality disorders. *International Journal of Eating Disorders*, *46*(8), 801–809.

Roberson, L., Kim, R., Russo, M., & Briganti, P. (2024). Please Excuse My Accent: An Examination of Impression Management Strategies Used by Nonnative Speakers. *Journal of Language and Social Psychology*, *43*(3), 298–325.

Salecha, A., Ireland, M. E., Subrahmanya, S., Sedoc, J., Ungar, L. H., & Eichstaedt, J. C. (2024). Large language models display human-like social desirability biases in Big Five personality surveys. *PNAS Nexus*, *3*(12), pgae533.

Sametoğlu, S., Pelt, D., Eichstaedt, J. C., Ungar, L. H., & Bartels, M. (2024). The value of social media language for the assessment of wellbeing: A systematic review and meta-analysis. *The Journal of Positive Psychology*, *19*(3), 471–489.

Saussure, F. de. (1962). *Cours de linguistique générale*. Payot.

Schaeffer, N. C., & Dykema, J. (2011). Questions for surveys: Current trends and future directions. *Public Opinion Quarterly*, *75*(5), 909–961.

Schwartz, H. A., Eichstaedt, J., Kern, M., Park, G., Sap, M., Stillwell, D., Kosinski, M., & Ungar, L. (2014). Towards assessing changes in degree of depression through facebook. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 118–125.

Schwartz, I. S., Link, K. E., Daneshjou, R., & Cortés-Penfield, N. (2024). Black box warning: Large language models and the future of infectious diseases consultation. *Clinical Infectious Diseases*, *78*(4), 860–866.

Sechrest, L. (1963). Incremental Validity: A Recommendation. *Educational and Psychological Measurement*, *23*(1), 153–158.

Shah, D., Schwartz, H. A., & Hovy, D. (2019). Predictive biases in natural language processing models: A conceptual framework and overview. *arXiv Preprint arXiv:1912.11078*.

Shankman, S. A., Funkhouser, C. J., Klein, D. N., Davila, J., Lerner, D., & Hee, D. (2018). Reliability and validity of severity dimensions of psychopathology assessed using the Structured Clinical Interview for DSM-5 (SCID). *International Journal of Methods in Psychiatric Research*, *27*(1), e1590.

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017). Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics*, *22*(5), 1589–1604.

Siegel, E. (2020). When does predictive technology become unethical. *Harvard Business Review*, *23*, 1–5.

Sikström, S., Nicolai, M., Ahrendt, J., Nevanlinna, S., & Stille, L. (2024). Language or rating scales based classifications of emotions: Computational analysis of language and alexithymia. *Npj Mental Health Research*, *3*(1), 37.

Sikström, S., Valavičiūtė, I., & Kajonius, P. (2025). Personality in just a few words: Assessment using natural language processing. *Personality and Individual Differences*, *238*, 113078.

Singh, H., Tiwari, S., Agarwal, S., Chandra, R., Sonbhadra, S. K., & Singh, V. (2025). Multimodal Data-Driven Classification of Mental Disorders: A Comprehensive Approach to Diagnosing Depression, Anxiety, and Schizophrenia. *arXiv Preprint arXiv:2502.03943*.

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S. F., & Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing*, *65*, 3-14.

Son, Y., Clouston, S. A., Kotov, R., Eichstaedt, J. C., Bromet, E. J., Luft, B. J., & Schwartz, H. A. (2023). World Trade Center responders in their own words: Predicting PTSD symptom trajectories with AI-based language analyses of interviews. *Psychological Medicine*, *53*(3), 918–926.

Soni, N., Matero, M., Balasubramanian, N., & Schwartz, H. A. (2022). Human language modeling. *arXiv preprint arXiv:2205.05128*.

Spitzer, R. L. (1983). Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry, 24*(5), 399–411.

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.

Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., ... & Bühner, M. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, *117*(30), 17680-17687.

Stefana, A., Damiani, S., Granziol, U., Provenzani, U., Solmi, M., Youngstrom, E. A., & Fusar-Poli, P. (2025). Psychological, psychiatric, and behavioral sciences measurement scales: Best practice guidelines for their development and validation. *Frontiers in Psychology*, *15*, 1494261.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*(1), 1–26.

Tackman, A. M., Sbarra, D. A., Carey, A. L., Donnellan, M. B., Horn, A. B., Holtzman, N. S., Edwards, T. S., Pennebaker, J. W., & Mehl, M. R. (2019). Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis. *Journal of Personality and Social Psychology*, *116*(5), 817.

Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., & Srikumar, V. (2016). A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of Substance Abuse Treatment*, *65*, 43–50.

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

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

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, *88*(3), 500.

Timmons, A. C., Duong, J. B., Simo Fiallo, N., Lee, T., Vo, H. P. Q., Ahle, M. W., Comer, J. S., Brewer, L. C., Frazier, S. L., & Chaspari, T. (2023). A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. *Perspectives on Psychological Science*, *18*(5), 1062–1096.

Tolin, D. F., Gilliam, C., Wootton, B. M., Bowe, W., Bragdon, L. B., Davis, E., Hannan, S. E., Steinman, S. A., Worden, B., & Hallion, L. S. (2018). Psychometric properties of a structured diagnostic interview for DSM-5 anxiety, mood, and obsessive-compulsive and related disorders. *Assessment*, *25*(1), 3–13.

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

Trimble, H. C., & Cronbach, L. J. (1943). A Practical Procedure for the Rigorous Interpretation of Test-Retest Scores in Terms of Pupil Growth. *The Journal of Educational Research*, *36*(7), 481–488.

Uher, J. (2013). Personality psychology: Lexical approaches, assessment methods, and trait concepts reveal only half of the story—Why it is time for a paradigm shift. *Integrative Psychological and Behavioral Science*, *47*, 1–55.

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

Varadarajan, V., Lahnala, A., Ganesan, A. V., Dey, G., Mangalik, S., Bucur, A.-M., Soni, N., Rao, R., Lanning, K., Vallejo, I., & others. (2024). Archetypes and entropy: Theory-driven extraction of evidence for suicide risk. *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, 278–291.

Varadarajan, V., Sikström, S., Kjell, O. N., & Schwartz, H. A. (2024). ALBA: Adaptive language-based assessments for mental health. *Proceedings of the Conference. Association for Computational Linguistics. North American Chapter. Meeting*, *2024*, 2466.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Vu, H., Nguyen, H. A., Ganesan, A. V., Juhng, S., Kjell, O. N., Sedoc, J., Kern, M. L., Boyd, R. L., Ungar, L., Schwartz, H. A., & others. (2024). PsychAdapter: Adapting LLM Transformers to Reflect Traits, Personality and Mental Health. *arXiv Preprint arXiv:2412.16882*.

Warrens, M. J. (2015). On Cronbach's alpha as the mean of all split-half reliabilities. *Quantitative Psychology Research: The 78th Annual Meeting of the Psychometric Society*, 293–300.

Warrier, U., Warrier, A., & Khandelwal, K. (2023). Ethical considerations in the use of artificial intelligence in mental health. *The Egyptian Journal of Neurology, Psychiatry and Neurosurgery*, *59*(1), 139.

Westermann, R., Spies, K., Stahl, G., & Hesse, F. W. (1996). Relative effectiveness and validity of mood induction procedures: A meta-analysis. *European Journal of Social Psychology*, *26*(4), 557–580.

Wille, N., Bettge, S., Ravens-Sieberer, U., & Group, B. S. (2008). Risk and protective factors for children's and adolescents' mental health: Results of the BELLA study. *European Child & Adolescent Psychiatry*, *17*, 133–147.

Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: Its limits, value, and potential for improvement. *Annu. Rev. Psychol.*, *55*(1), 493–518.

Wong, A., Young, A. T., Liang, A. S., Gonzales, R., Douglas, V. C., & Hadley, D. (2018). Development and validation of an electronic health record–based machine learning model to estimate delirium risk in newly hospitalized patients without known cognitive impairment. *JAMA Network Open*, *1*(4), e181018–e181018.

Wright, A. J., Pade, H., Gottfried, E. D., Arbisi, P. A., McCord, D. M., & Wygant, D. B. (2022). Evidence-based clinical psychological assessment (EBCPA): Review of current state of the literature and best practices. *Professional Psychology: Research and Practice*, *53*(4), 372–386.

Wright-Berryman, J., Cohen, J., Haq, A., Black, D. P., & Pease, J. L. (2023). Virtually screening adults for depression, anxiety, and suicide risk using machine learning and language from an open-ended interview. *Frontiers in Psychiatry*, *14*, 1143175.

Wynn, C. J., Barrett, T. S., & Borrie, S. A. (2024). Conversational speech behaviors are context dependent. *Journal of Speech, Language, and Hearing Research*, *67*(5), 1360–1369.

Xintong, Z., & Xiaofei, H. (2024). The Relationship Between Linguistic Features and Psychological States: A Quantitative Approach. *Medical Research Archives*, *12*(8).

Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., Ghassemi, M., Dey, A. K., & Wang, D. (2024). Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *8*(1), 1–32.

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, *55*(1), 90–112.

Yap, A. U., Sultana, R., & Natu, V. P. (2024). Somatic and temporomandibular disorder symptoms-Idioms of psychological distress in Southeast Asian youths. *CRANIO®*, *42*(4), 364-371.

Yates, A., Cohan, A., & Goharian, N. (2017). Depression and self-harm risk assessment in online forums. *arXiv Preprint arXiv:1709.01848*.

Yawer, B. A., Liss, J., & Berisha, V. (2023). Reliability and validity of a widely-available AI tool for assessment of stress based on speech. *Scientific Reports*, *13*(1), 20224.

Yin, H., Zhu, H., Gu, J., Qin, H., Ding, W., Guo, N., Fu, J., & Yang, Y. (2024). Mobile-based ecological momentary assessment and intervention: Bibliometric analysis. *Frontiers in Psychiatry*, *15*, 1300739.

Yonashiro-Cho, J. M., Gassoumis, Z. D., Wilber, K. H., & Homeier, D. C. (2021). Improving forensics: Characterizing injuries among community-dwelling physically abused older adults. *Journal of the American Geriatrics Society*, *69*(8), 2252–2261.

Youngstrom, E. A., Choukas-Bradley, S., Calhoun, C. D., & Jensen-Doss, A. (2015). Clinical guide to the evidence-based assessment approach to diagnosis and treatment. *Cognitive and Behavioral Practice*, *22*(1), 20–35.

Youngstrom, E. A., Van Meter, A., Frazier, T. W., Hunsley, J., Prinstein, M. J., Ong, M.-L., & Youngstrom, J. K. (2017). Evidence-based assessment as an integrative model for applying psychological science to guide the voyage of treatment. *Clinical Psychology: Science and Practice*, *24*(4), 331–363.

Zhao, X., Feng, G. C., Ao, S. H., & Liu, P. L. (2022). Interrater reliability estimators tested against true interrater reliabilities. *BMC Medical Research Methodology*, *22*(1), 232.

Printed by Media-Tryck, Lund 2025 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

Faculty of Social Sciences
Department of Psychology

ISBN 978-91-8104-705-9