LUNDS UNIVERSITET

Department of Psychology

The ALIRT Model:

An Adaptive Language-Based Assessment Model for Diagnosing Mental Disorders

Rebecca Astrid Böhme

Master's Thesis
2024

Supervisor: Sverker Sikström

Abstract

In this project, I investigate the combination of Natural Language Processing (NLP) with Item Response Theory (IRT) to advance the assessment and scoring of open response items. Traditional psychometric assessments are grounded in well-defined quality standards, yet assessments based on natural language have missed meeting these standards. To address this gap, I integrate NLP processed open response items in an IRT framework, aiming to combine the strengths of natural language processing and item response theory to enhance mental health assessment and establish a foundation for computerized adaptive testing.

In this study, I address three central research questions: the adequacy of newly developed open response items in capturing DSM-5 criteria for initial mental health assessments, the accuracy and efficiency of the ALIRT model in diagnosing common mental disorders, and the improvement in validity when combining open response items with traditional rating scales. I hypothesize that the ALIRT model will provide accurate and valid initial diagnoses. It will require fewer questions than traditional methods and show higher accuracy in categorizing mental health disorders through open-response items. I further hypothesize that it offers greater ecological validity and reduced diagnostic time, and that open-ended responses will be preferred over traditional rating scales.

The findings, while limited, offer valuable insights for the future development of this approach. The model comparison indicates that the mixed model is superior in the current modeling approach. However, I discuss several limitations encountered during the study, including the complexities of integrating open responses into an IRT framework. Future work will focus on addressing the identified limitations, refining the model, and exploring additional applications of this approach in computerized adaptive mental health assessments.

Keywords: mental health, assessment, artificial intelligence, natural language processing, item response theory

Introduction

The description and understanding of psychological phenomena, and thus the individual's experience are at the core of mental health assessments. In recent years, the field of psychology developed an increased awareness of the simplifications and limitations present in traditional mental health assessments. These metrics mainly rely on fixed-response rating scales that do not fully capture the complexity and nuances of individual experiences. With the development of new technologies, such as natural language processing (NLP), the inspiration to integrate these modern advances in existing psychometric frameworks sparked. Instead of restrictedly responding on a numerical scale, individuals would get the opportunity to use their own words to describe their inner state and experience. This integration has the potential to revolutionize the field by creating more adaptive, nuanced, and individualized diagnostic tools. An especially beneficial development in settings, where specialized mental health assessments are limited. For example, in Sweden, most patients with mental health issues are initially seen by General Practitioners, who may not provide the same level of specialized assessment as mental health professionals (Sundquist et al., 2017). Utilizing these novel techniques to inform AI-based digital health solutions could support assessments by offering additional, nuanced information, ultimately benefiting both clinicians and patients.

In this project, I focus on the development of a model that has the potential to inform such an AI-based digital health solution. This project builds on previous work in the research group, where a pioneering adaptive language-based assessment model was developed that integrates NLP in an IRT framework (Varadarajan et al., 2023). In this study, I expand the scope of the previous model to focus on a wider range of mental disorders, moving beyond depression and anxiety assessment, and including bipolar disorder, obsessive-compulsive disorder, eating disorders, attention-deficit hyperactivity disorder, autism spectrum disorder, substance abuse disorder, and post-traumatic stress disorder.

Prevalence and Impact of Mental Health Problems

We experience a high prevalence of mental health problems globally (Steel et al., 2014). For example, mood disorders such as depression or anxiety have been identified as the main cause of disability (Vos et al., 2016). These developments, combined with a reduction of stigma, lead to an increased amount of people seeking help and mental health support. Additionally, we experience a shortage of staff that can diagnose those in need, guide them, and provide them with the help they need (Butryn et al., 2017; Harahan, 2010). The World Mental Health Report describes this globally present problem and states that mental health

services are ill-equipped (WHO, 2022). A structured mental health assessment is the key to a reliable identification and diagnosis of mental disorders. It guides treatments and monitors progress. Traditionally, psychometric assessments have relied on standardized rating scales (SRS) and structured interviews, grounded in established quality standards. Structured interviews, while invaluable, demand a well-trained professional and time resources that are not given in every healthcare context. SRS, while effective, can be limited by their rigidity and inability to capture the nuanced, personalized responses of individuals, which will be discussed in more detail later. In a natural setting, individuals communicate with language, and, indeed, previous research has shown that open responses are preferred due to its option to expand and precisely describe the individual experience (Sikström, Höök, et al., 2023). The utilization of Artificial Intelligence (AI) provides novel solutions to assess individuals efficiently and is recognized as a toolbox to optimize patient care (Mittal et al., 2023).

Item Response Theory

Item response theory (IRT) is a measurement paradigm commonly used in the development of scales and tests. In recent years, IRT is increasingly used as a framework to test theories, and to better understand human reactions and behavior (Lang & Tay, 2021). IRT offers the opportunity to examine individual reactions on an item level and model interindividual variations (Edelen & Reeve, 2007). In this framework, the assessed item parameters and item characteristics model the person's response and ability score (Edelen & Reeve, 2007). For example, difficulty parameter indicates how challenging an item is, which helps in distinguishing between different levels of ability. The discrimination parameter indicates how well an item can differentiate between individuals with varying levels of the trait being measured. The guessing parameter is introduced to quantify a participant's random selection of an item.

In relation to my research question, IRT gives the opportunity to assess different symptoms and behavioral expressions of mental disorders by examining the individual's response to the presented questions. One can quantify the individual's latent traits or abilities, such as the severity of specific symptoms or the presence of certain behavioral patterns. This quantification allows to create a detailed profile of the individual's mental health status, and is used in clinical practice to identify areas that may require intervention, and monitor changes over time. Additionally, IRT offers an understanding of how different items on a test contribute to the overall measurement of the disorder, ensuring that the assessment is both accurate and reliable. For example, to describe the underlying structure, one differentiates between unidimensional and multidimensional IRT. While unidimensional IRT assumes one

underlying construct, multidimensional IRT assumes multiple constructs. A unidimensional assumption is in many cases not viable since most items are likely to explain variance of different constructs (Kim & Lee, 2023). The same applies to the assumption of local independence, which describes the condition where, given a respondent's level of the latent trait, the responses to individual items on a test are statistically independent of each other (Kim & Lee, 2023). A violation of these assumptions, speaks for a multidimensional structure. Considering the great overlap of mental health symptoms, assuming a multidimensionality for mental disorders seems reasonable on a theoretical level as well.

In the development of SRS, these considerations and assumptions are applied. SRS are developed based on the observation that certain behaviors or feelings underlie a common cause and therefore, represent one common latent. In an IRT context the terms ‘ability’, ‘latent trait’, ‘latent’, or ‘factor’ are used interchangeably, and reference to the underlying construct that is assessed through the items. In the context of mental health, this cause (latent) is often categorized as a diagnosis, which is a summation of symptoms reflecting an alteration of thoughts, feelings, and consequently behavior. To make these alterations quantifiable and accessible, SRS have been successfully used for decades. SRS are combining a variety of items which aim to capture the greatest amount of variance possible to describe one construct. In summation, these items represent the underlying cause (i.e., latent).

IRT was continuously developed, introducing a variety of standards to ensure the highest quality possible. This is a great advantage. While developing novel rating scales, researchers follow the criteria established in the field, which assures the same quality for novel scales. An advantage is that for trained practitioners SRS are easy and quickly applicable. However, SRS come with some drawbacks. For example, SRS are rather rigid and restricted. Psychological constructs are clearly defined and theoretically distinct, but when assessing them with SRS we observe that diagnostic criteria share a great amount of variance and are not as distinguishable as theorized. These, so called, cross-loadings reduce the information and predictive power that can be extracted from a single construct. Furthermore, SRS are often rather long, including many items. While item reduction in established psychometrics is beneficial to describe the construct of interest more accurately and reduce response fatigue (Haroz et al., 2020), it also requires big samples with many data points. Computerized Adaptive Testing (CAT) is a testing technique developed to solve this disadvantage. This computational solution allows to reduce the set of test items to an optimal amount and, simultaneously, increases testing precision. CAT selects the optimal next item based on the highest informational value in comparison to the previous one by utilizing

optimization algorithms (Chalmers, 2016). In this way, an optimized and reduced set of items is presented compared to fixed test batteries. CAT is the most optimized state-of-the-art assessment present in the field of psychological assessment.

Natural Language Processing

The AI field is rapidly growing and includes a variety of approaches. AI technology can be found in the daily lives of everyone and is extremely diverse. Natural Language Processing (NLP), as a form of AI, uses machine learning algorithms to enable the processing and quantification of natural language computationally (Khurana et al., 2023). Clearly, the recent developments in NLP are of great interest in fields, where the aim is to capture and quantify individuals' inner state, thoughts, and feelings. NLP follows the general idea, that language follows a set of symbols and rules, and can be understood as such (Khurana et al., 2023). In this approach, one assigns probabilities to rules and symbols to predict expected appearances. This logic allows us to observe specific language patterns used to assess various language features. Two main approaches are introduced: decontextualized and contextualized word embeddings. Decontextualized word embeddings (e.g., word2vec, GloVe) do not consider context, while contextualized word embeddings (e.g., BERT) incorporate the context in which words appear. The development of the Bidirectional Encoder Representations from Transformers (BERT) model in 2018 marked a significant breakthrough in using contextualized word embeddings efficiently. This advancement has enabled NLP to capture contextual nuances, making the BERT model powerful and suitable for diverse applications, such as diagnostics through open-ended questions or clinical interviews.

When processing language, infinite language representations are reduced to a finite set, while assuming that language has less ambiguity than in reality (Chowdhary, 2020). While this a computationally necessary assumption, it shows a challenge within language processing. With the development of BERT, some aspects of ambiguity were resolved, thus a more precise, contextual language representation is possible. This progress in the development significantly increased the efficiency and precision of language analysis by providing methods beyond linear modeling. NLP provides an automated decision on language interpretation over domains (Glaz et al., 2021).

In the past years, AI research has focused on different domains within the mental health sector and was increasingly used to help diagnose patients (e.g., Aditya Shastry & Sanjay, 2023; Gagliardi et al., 2021; Kishimoto et al., 2020). The overall goal of utilizing AI in health care is to support practitioners and make their lives easier by outsourcing tasks that

can be done automatically. This will save time that can be invested in a context where human-human interactions are crucial. For example, when practitioners need time with the patient to understand their complex symptoms and individual needs to provide personalized medical care and treatment options. Here, AI-based solutions hold the potential to indirectly increase the quality of several aspects of mental health care by providing more time.

AI has also become of interest in different kinds of assessments, especially in areas where usually interviews or SRS are used. In this assessment field, an automated, objective, structured, and validated way to quantify qualitative data is still missing. This task would be usually performed by professionals, who are trained over a long period to reliably hand-code interviews or use SRS to assess participants. Here, NLP becomes particularly interesting since the currently used approaches are not only time-consuming but also very costly, furthermore, it has been criticized as potentially biased (Levitt, 2021).

NLP has been considered a new paradigm in clinical research (Glaz et al., 2021). In some areas, we see rapid developments and implementations of this novel technique. In the neuropsychological field, for example, AI-based tools are increasingly used to help with the assessment of cognitive decline (Graham et al., 2020; Moret-Tatay et al., 2021). Aging alters language processing, which can be analyzed and provides linguistic markers. Several studies have shown that these alterations are reliably identified by AI-based tools (Agbavor & Liang, 2022; Beltrami et al., 2018; Broderick et al., 2021; Diaz-Asper et al., 2022; Fraser et al., 2019; Penfold et al., 2022; Yeung et al., 2021). Furthermore, these results hold in different languages (e.g., Igarashi & Nihei, 2022; Metarugcheep et al., 2022). In other areas, as in general mental health assessment, there is less implementation and a greater potential for further developments.

Limitations of Standardized Rating Scales

Standardized rating scales (SRS) have a valuable place in mental health assessment. However, the closed-ended format comes with certain limitations. Some research suggests that SRS overlook the presence of other, equally important symptoms (Glaz et al., 2021). For example, depression scales show a lack of symptom overlap, suggesting a higher syndrome complexity (Fried, 2017). This supports the benefit of natural language processing (NLP) to assess mental health and further define present mental health constructs to understand interconnected and complex symptoms. Previous studies have shown that open-ended responses can predict corresponding SRS outcomes (Fatima et al., 2021; K. Kjell et al., 2021; O. N. E. Kjell et al., 2019; Li et al., 2020), justifying the use of open responses as a valid assessment format.

One significant limitation of SRS is that they constrain participants to predefined answers, preventing them from freely describing their subjective experiences and individual states. Recent research has demonstrated that respondents experience open-ended responses in free text as more precise (Sikström, Höök, et al., 2023), and these responses show competitive or higher validity and reliability compared to traditional rating scales (O. N. E. Kjell et al., 2019). Additionally, open-ended responses provide higher ecological validity, which is a notable advantage over SRS. By quantifying text based on machine learning techniques, we can reduce potential biases in the diagnostic setting, offering a comprehensive understanding of the respondent's mental health.

Moreover, standardized rating scales often fail to capture the complexity of mental health symptoms. For example, two individuals with the same score on a depression scale will have different experiences of their depression, with one experiencing severe anhedonia and the other suffering from intense feelings of worthlessness. SRS do not account for such variations, potentially leading to oversimplified assessments and consequently, interventions. In contrast, open-ended responses allow for a richer, more detailed depiction of the respondent's mental state, which can be crucial for accurate diagnosis and personalized treatment planning.

SRS can be subject to various biases, including social desirability bias and response style bias, where individuals might consistently choose moderate responses and avoid extreme options. Open-ended responses can reduce these biases by focusing on the content and sentiment of the text rather than the fixed options provided in a scale.

Natural Language Processing as Complementing Technique

NLP is not yet an established approach in the assessment of mental health. Considering the usage of NLP, we miss the focus on the assessment of clinically relevant mental health constructs (Ahmad et al., 2020; Mittal et al., 2023). To utilize these techniques in clinical practice, we need to focus on the topics that are mainly relevant in diagnostic areas. This research group has previously focused on such constructs (e.g., K. Kjell et al., 2021; O. N. E. Kjell et al., 2019; Sikström et al., 2023).

Another crucial point is that we lack important quality standards to assure reliability, validity, and objectivity when using NLP. While SRS are evaluated based on well-defined quality standards, the evaluation of NLP models is poor and not standardized. We need techniques to achieve the same evaluation and apply the same high-quality standards. This is a crucial step to assure patient security, a fair assessment, and following adequate treatment.

Rationale of the Study

Despite significant advancements in natural language processing (NLP), its application in mental health assessments remains underdeveloped. Traditional psychometric methods lack the flexibility to adapt to the unique ways individuals express their mental health experiences. Conversely, NLP has shown promise in analyzing open-ended responses but lacks the quality standards necessary for reliable assessments. To address these identified gaps, this study proposes the integration of NLP within an IRT framework to enhance mental health assessments. In this project, I combine the best parts of SRS and NLP to achieve the highest quality standard possible when assessing mental health constructs.

The project holds a variety of potential benefits. The underlying IRT structure allows the presentation of the most informative questions to the participants and reduces the assessment time. The utilization of NLP allows the individuals to freely describe their inner states and experiences. The computerized adaptive testing (CAT) structure facilitates a fully automated initial diagnosis which counteracts the increasing shortage of professionals within the field of psychology. Furthermore, the proposed approach provides a more objective assessment than classical clinical interviews and overcomes biases. For example, the relationship between practitioner and patients is a well-investigated, driving effect for therapeutic success (e.g., Del Re et al., 2021; Høglend, 2014). During a first contact, a mismatch between practitioner and patient might hinder an appropriate assessment. Discussing openly in a session for an initial diagnosis might be more difficult for some individuals when they do not feel safe with the person they are talking to. At this moment face-to-face consultations become a disadvantage. Here, automated processing has the potential to reach patients that might hold back and allow them to express themselves without fear. Furthermore, the proposed approach allows an assessment beyond geographical barriers and will reach individuals who have no or complicated access to mental services.

In summary, the high standards of IRT-based developments and computerized adaptive testing, combined with NLP's potential to analyze complex language data, offers a novel and potentially highly beneficial approach in capturing mental health and supporting clinicians with scarce resources.

Aim of the Study

The aim of the study is to combine advantageous aspects of natural language processing (NLP) and item response theory (IRT) to assess mental health and to provide the basis for computerized adaptive testing. While traditional psychometrics are developed based on clearly defined quality standards (e.g., reliability, validity, and objectivity measurements),

assessments based on natural language still lack those. To close this gap, I bring the best parts of both approaches together by applying the IRT to open-ended questions.

In this study, I will address 3 central research questions:

(1) What DSM-5 criteria are commonly used to assess mental health during the initial contact and can they be sufficiently captured by the developed open response items?
(2) How accurate is the ALIRT model in providing an initial diagnosis for common mental disorders, and how many questions does it require compared to traditional methods (i.e., SRS)?
(3) Does the combination of open response items with traditional rating scale items improve the validity of mental disorder categorization compared to one single approach (i.e., SRS vs. OEQ)?

Hypotheses

The main research question investigates whether the combination of IRT and NLP allows an optimized diagnostic of the mental disorders of interest that can replace or complement SRS. I hypothesize that the ALIRT model will provide an accurate and valid initial diagnosis for the defined mental health disorders (H1). I hypothesize that the ALIRT model will need a reduced number of questions to assess the mental health disorder of interest, compared to SRS (H2). I hypothesize that open-ended questions show a higher accuracy in categorizing mental health disorders than SRS (H3). I hypothesize that the ALIRT model will provide more ecological validity and a reduced diagnostic time (H4). Considering previous findings, I hypothesize that open-responses are preferred compared to SRS (H5).

Methods

This study is embedded in a larger project, focusing on AI-based language models to assess mental health. The project is funded by the '*Marianne och Marcus Wallenbergs Stiftelse*' (Project ID: MMW-2021.0058) and ethical approval is provided within this project (2024-00378-02).

Within this study sensitive, but no personal data was collected. I followed the general recommendations by the Swedish Ethical Review Authority and the research is in line with the Declaration of Helsinki. To assure data security, all data will be handled with great care, strictly following the guidelines of the General Data Protection Regulations (GDPR). Prior to participants' involvement in the study, I informed them about the research objectives and procedures. Participants' provided written consent prior to the data collection, and after they received all necessary information. Participants had the possibility to withdraw from the study at any time, without provided explanation or facing any consequences. I informed the participants about the possibility to experience some discomfort when reporting and writing about their mental health and referred to possible support offers in case they are needed. All information was provided in English.

Design

This validation study concentrates on the development and validation of the Adaptive Language-based Item Response Theory (ALIRT) model. I divided the study into two main phases, where in the first phase I developed stimulus material (i.e., open response items) and pre-screened the participant pool to inform Phase 2. In Phase 2, I collected mental health data in the form of SRS, participant's free-text narratives, and descriptive words. I focus on the 9 common mental disorders including mood disorders (i.e., depression, anxiety, bipolar disorder) as well as substance use disorder (i.e., alcohol and/ or drug abuse), autism spectrum disorder, attention deficit hyperactivity disorder/ attention deficit disorder, eating disorders, obsessive-compulsive disorder, and post-traumatic stress disorder.

Participants

I collected data from 550 participants electronically over the online platform Prolific (https://www.prolific.co) and compensated each participant with £6.29/hr to £9.84/hr, depending on the time used to complete the study. For each diagnosis, 50 participants were recruited (total = 450) along with 100 healthy control participants. The number of participants was determined through power analysis performed by the collaborator based on the previously conducted study (Varadarajan et al., 2023). In a prior data collection, I prescreened the participants to ensure that the participants had an ongoing diagnosis (i.e.,

major depression disorder, generalized anxiety disorder, bipolar disorder, obsessive compulsive disorder, attention-deficit/hyperactivity disorder, autism spectrum disorder, eating disorder, substance abuse disorder, post-traumatic stress disorder), assessed the treatment status, and whether this diagnosis was given by a professional or not. I included participants with multiple diagnoses to provide a higher ecological validity during the model training. Furthermore, participants had to be 18 years or older, live in the USA, and have English as their first language. I excluded participants when they missed at least one of the four attention checks, furthermore, participants were excluded when their responses included only non-word characters, repeated characters, or single characters. The final sample includes N = 515 participants (297 female, 35 non-binary, 1 preferred not to answer) with a mean age of 38.89 years (*SD* = 12.26, *range* 18 to 78).

Procedure

Phase 1

In Phase 1, the author with support from the research group developed open response items to assess participant's mental health where most response alternatives consisted of 2 to 5 words in addition to narratives. In some cases, additional rating scale items were added to assess the binary diagnostic criteria. Here, each question intends to assess either crucial symptoms, frequency, or onset of the symptoms to capture the participants' full experience. The stimulus material was developed through multiple steps. First, the author of this thesis developed questions based on the DSM-5 criteria. The questions aim for the highest clarity possible, which is why some of the questions are represented as traditional ratings. Second, the questions have been independently and qualitatively evaluated by 3 clinical psychologists out of this research group and one external clinical psychologist who specializes in diagnosing ADHD and ASD. Suggested changes have been discussed and implemented leading to the final set of 50 questions. I aimed to ensure the consistency and logic of the captured information, furthermore, through active stakeholder engagement, I ensured the best representation of questions possible. Third, the questions were evaluated by the collaborating computer scientist and methodological expert of the group. Here, we discussed statistical implications to aim for a balance between latent representation and captured variance.

Phase 2

In Phase 2, I invited the pre-screened participants to take part in the study. The participants were informed about the general style of the study and the compensation beforehand. Furthermore, they were informed about the right to withdraw their participation at any time and gave written consent. Due to the sensitive nature of the data collection,

participants were informed about the possible risk of discomfort and have been provided with possible help offers in case they are needed. I informed participants that the entire data collection was anonymous, and that no personal identification would be possible. Furthermore, the participants were informed about the data processing and storage, which is in accordance with the General Data Protection Regulation (GDPR). From the initial recruitment platform *Prolific*, participants were transferred automatically to *Qualtrics* where all the data was collected. Here, the participants answered 3 blocks. The first block included the developed questions from Phase 1, the second block included the SRS, and the third block included screening questions (e.g., treatment and comorbidities). The questions in the first block, as well as the rating scales in the second block, have been randomized to avoid ordering effects. However, to obtain the intended structure of the rating scales the individual questions within one scale were not randomized. Participants had no time restriction to complete the study. The median time in the groups ranged from 1h00min to 1h35min.

Material

Open-ended Questions

The developed material includes 50 questions. Two questions are essay questions, capturing the participant's narratives focusing on general mental health and traumatic events. Here, participants have been asked to provide a paragraph with at least 300 characters. Furthermore, I included three questions asking for a binary response and one question including a traditional categorical response. One question asks for substances used. The other 44 questions are open-ended questions providing the participant with the possibility to provide 1 to 5 words to describe their mental state. The developed questions are categorized based on their content, while some of them capture diagnosis-specific symptoms, others are considered general questions and address symptoms which tap into different diagnoses.

Traditional Rating Scales

I used 10 established rating scales to assess the mental health constructs of interest. All questionnaires are openly accessible and regularly used to diagnose the respective mental disorders. The Patient Health Questionnaire-9 (PHQ9; Kroenke et al., 2001) was used to assess depression. This questionnaire uses 9 items to assess depressive symptoms on a 4-point Likert scale. I used the General Anxiety Disorder-7 Scale (GAD-7; Johnson et al., 2019) to assess anxiety. This questionnaire uses 7 items on a 4-point Likert scale. I used the Mood Disorder Questionnaire (MDQ; Hirschfeld et al., 2000) to assess Bipolar Disorder. This questionnaire includes 14 items with a binary response (Yes/No) and one item with a 4-point Likert scale. I used two different questionnaires to assess substance abuse. The Alcohol

Use Disorder Identification Test (AUDIT; Dawson et al., 2005) was used to assess alcohol abuse. The AUDIT uses 8 questions with a 5-point Likert scale and 2 questions with a 3-point Likert scale. The Drug Abuse Screening Test (DUDIT; Hildebrand, 2015) was used to assess drug abuse. Here, 9 questions with a 5-point Likert scale and 2 with a 3-point Likert scale are used. I used Part A of the Adult ADHD Self-Report Scale (ASRS; Kessler et al., 2005) to assess Attention-Deficit/Hyperactivity Disorder. This questionnaire uses 6 questions with a 5-point Likert scale to screen for ADHD. Part B does further elaborate on the concrete symptoms the patient experiences but can be disregarded for the screening. The Ritvo Autism and Asperger Diagnostic Scale (RAADS-14; Eriksson et al., 2013) was used to assess Autism Spectrum Disorder. This questionnaire uses 14 items with a 4-point Likert scale. I used the National Stressful Events Survey PTSD Short Scale (NSESSS-PTSD; LeBeau et al., 2014) to assess Post-traumatic Stress Disorder. Here, one open text response (i.e., keyword to capture the traumatic event) and 9 items on a 5-point Likert scale are used. The Brief Obsessive–Compulsive Scale (BOCS; Bejerot et al., 2014; Patel et al., 2022) was used to assess Obsessive-Compulsive Disorder. The BOCS uses 15 items on a 3-point Likert scale and one open-response question (i.e., separates the percentage of obsessions and compulsions). I used the Eating Disorder Examination Questionnaire (EDE-QS; Gideon et al., 2018; Prnjak et al., 2020) to assess eating disorders. Here, 12 items on a 4-point Likert scale are used.

Analysis

In this study, I employ unidimensional and multidimensional latent trait models in an Item Response Theory framework to analyze, both, dichotomous and polytomous response data. All analyses were performed in R (version 4.3.2.) and Python (3.11.7). The code and data are available to the evaluators of the thesis in the following GitHub repository (https://github.com/rebeccaboe/MSc-ALIRT). The data analyses will be explained in 5 different stages: Preprocessing, Polytomization, Item Response Modeling, Computerized Adaptive Testing, and Evaluation.

Preprocessing

The data preprocessing included data cleaning. As previously mentioned, I excluded all participants who failed at least one attention check or provided open responses of low quality. This included responses that were irrelevant, nonsensical, or did not address the questions asked. Examples of such responses included random strings of characters, unrelated statements, or answers that did not pertain to the context of the mental health assessment. During the preprocessing, I identified the participant's prescreened diagnostic groups and labeled them accordingly. The open responses were converted to lowercase, punctuation was

removed, and the responses were tokenized, which are common procedures in NLP. Open responses are best represented as word embeddings, which is why I used a Bidirectional Encoder Representations from Transformers (BERT) model, specifically '*bert-base-uncased*'. The special feature of these models is that they allow a contextual representation, which goes beyond word counts in language representation. I replaced missing words with the token '*UNK*', a placeholder was necessary since the participants had the option to provide between 1 to 5 descriptive words. During training, the presence of the '*UNK*' token helps the model to handle unseen or missing words and ensures the generation of meaningful output. I re-coded the reversed items in the SRS and calculated scale scores as stated in the respective manual of each scale.

Polytomization

A polytomization, the conversion of a continuous scale into a categorical representation, allows data simplification and brings the responses into a format compatible with an IRT paradigm. The quantification of natural language results in word embeddings, which are high dimensional vector representations (i.e., 768 vectors for each embedding). The primary objective of the polytomization step was to reduce the dimensionality but preserve as much informational value as possible of the embeddings. Another advantage of this step is that it also allows the usage of existing IRT packages (e.g., in R or Python) to facilitate comprehensive analysis. I utilized a prediction model as the basis for the polytomization. The developed linear regression-based prediction model uses word embeddings and generates graded responses at intervals of 25%, 50%, 75%, and 100%, indicating the probability that the provided response informed the correct diagnosis. Here, the correct diagnosis was informed by the pre-screened groups. I implemented 10% leave-one-out cross-validation to reduce overfitting and ensure the robustness and predictive accuracy of the model. This technique systematically leaves out 10% of the data for validation while training the model on the remaining 90%, repeating this process until each subset has been used for validation. By doing so, I ensure that the model's performance is tested on all data points, providing a comprehensive assessment of its generalizability and reliability across different subsets of the dataset. The chosen methodology allowed me also to determine the diagnostic accuracy of open responses, narratives, and rating scales by using the assigned diagnostic groups as the ground truth. Here, I use a multiclass classification algorithm to determine the truely predicted instances. Additionally, I utilized a logistic regression-based prediction model to generate binary responses, indicating whether the provided input

informed the diagnosis. This binary format enabled me to effectively compare the responses to the MDQ, which also has binary outcomes.

Item Response Modeling

IRT, as previously described, is the measurement framework commonly used for the development of rating scales, and includes a compilation of different analytical approaches to determine interindividual differences on item level (Edelen & Reeve, 2007). The item response modeling informs the specifications used within this framework to define the model.

As an initial step, I developed an item bank. An item bank is the foundation for adaptive testing, providing a pool of calibrated items to inform the assessment. In this step, I considered the data based on the construct of interest (e.g., depression, anxiety, eating disorder etc) and ran assumption checks first. I generated correlation plots and investigated the intercorrelations between the items to assure a robust relationship and avoid multicollinearity. I examined a screeplot to verify the unidimensional structure of the data. After that, I ran unidimensional IRT models for every rating scale and for every set of open response items.

The IRT framework offers a variety of models (e.g., 2PL, Graded Response Model, or Partial Credit Model) to understand the relationship between item response and ability level. Depending on the item structure (i.e., dichotomous, polytomous) and the item parameters of interest (e.g., discrimination, difficulty, or guessing), the specific model is selected. I selected a Graded Response Model (GRM), since it is a suitable model for item scorings which are reflecting an increased presence of the construct of interest (Chalmers, 2016). As the gradings are obtained through a prediction model establishing probabilistic segments for the open response items, this appeared to be the best estimation choice. I did not introduce any guessing parameter due to the nature of the questions (i.e., health-related data) where I assume that items can not be randomly answered or guessed, as it would be the case in learning or educational assessments. I used the Monte Carlo Expectation Maximization (MCEM) algorithm for estimating the parameters. This estimator was selected since it combines the Expectation Maximization algorithm with Monte Carlo simulation, which makes it particularly suitable for multidimensional data (Chalmers, 2016; Luo, 2018).

I evaluated the item properties of the IRT models to determine items with good properties. These properties are indicators for how well the item represents the underlying construct. The main focus was the variance explained by the item and how well it discriminates along the latent (i.e., along the ability scores). I disregarded items with discrimination parameter (α) below 1, as it is common practice in IRT modeling (Embretson

& Reise, 2000; Lai et al., 2003). I, furthermore, evaluated person and item infit and outfit statistics. The squared mean of the item infit is the ratio between predicted and observed variance, where 1 is described as ideal value (Lai et al., 2003). However, for item fit statistics, one considers values between 0.5 and 1.5 as productive for measurement and items within this range are used for measurement (Chalmers, 2016; Edelen & Reeve, 2007). The person fit indices allow an evaluation on how consistent the observed response patterns are with the proposed model, where z-values between -2 and 2 are preferable (Edelen & Reeve, 2007; Embretson & Reise, 2000). Furthermore, I inspected the Item Characteristic Curve (ICC) and Item Information Curve (IIC) for each item, ensuring a reasonable representation of the ability level (theta) and the level of information. The Test Information Curve was used to evaluate the summarized performance of the item compilation. For a comprehensive overview of the IRT models, item properties, and graphic representations, follow the designated path in the respective GitHub repository.

After developing the item bank, I explored the data construct-independent to determine if the presented latent structure is consistent with the theorized diagnostic categories. In respect to the research questions, I performed the analyses for 3 data compositions. I considered either the rating scale items, the NLP-processed open response items (i.e., word embeddings), or the combination of the rating scale item and NLP-processed open response items.

The data exploration aimed to inform the IRT model. That means that for each data composition, I explored how the data is best represented considering the data structure and the theoretical basis. I began the data exploration by generating a heatmap to visualize the intercorrelations among the variables, allowing me to identify patterns and relationships within the dataset. I conducted an Exploratory Factor Analysis (EFA) to uncover the underlying latent structure and to better understand the dimensions represented in the data. As before, I used a screeplot displaying the Eigenvalues and cumulative variance to determine the number of latents. The latents were extracted based on Eigenvalues > 1 (Goretzko et al., 2021; Watkins, 2018). Additionally, I evaluated factor loadings and fit indices to determine the latent structure of the data. Here, I set the criteria to loadings (λ) $\geq$ 0.30 as an indicator of the relationship between the item and latent (Goretzko et al., 2021). In the next step, I informed the IRT model with the EFA results. Based on the represented structure, I run a multidimensional IRT model. During this item response modeling the same model specifications as for the development of the item bank are used.

To evaluate the IRT models, I used a variety of criteria. As before, I evaluated the item parameter, focusing on the discrimination parameter (a) indicating how sensitive an item is to differences in the latent trait (θ). Here, higher values suggest that the item is more effective at distinguishing between responses (Edelen & Reeve, 2007; Keetharuth et al., 2021). I evaluated the general model fit by inspecting the Comparative Fit Index (CFI) and the Tucker Lewis Index (TLI), where values > 0.95 indicate a good model fit (Hu & Bentler, 1999). Additionally, I examined the root mean square error of approximation (RMSEA), with a cutoff < .05 indicating close fit and < .08 indicating reasonable fit, and standardized root mean square residual (SRMR), with a cutoff < .08 indicating good fit and < .1 indicating reasonable fit (Hu & Bentler, 1998, 1999; Kaiser, 1974). Furthermore, I inspected person fit and item fit parameters and I visually inspected the item representations. I plotted Item Characteristics Curves to graphically evaluate the item's discrimination over the ability. To judge the level of information captured by the item, I plotted Item Information Curves. To evaluate the item compilation, I plotted a Test Information Curve and a Scale Characteristic Curve. As a final evaluation step, I evaluated if the model is valid and appropriately calibrated by comparing the latent IRT score with the simple number-correct-score (Chalmers, 2016).

Computerized Adaptive Testing

The ALIRT project aims for computerized adaptive testing (CAT) comparing and combining rating scale responses and open response items. Here, I follow a similar approach that has been successfully used in an earlier study by this research group (Varadarajan et al., 2023), with the extension to a wider selection of mental disorders. To implement the computerized adaptive testing in R, I follow six different steps as defined by Erdem Kara (2019). Here, I define the IRT model; inform the item pool; define a starting rule; select an item selection rule; select a scoring rule; and lastly, I define the termination criteria. The first two steps are informed by the previously explained IRT analysis. The model specifications for the CAT are consistent with the previously estimated models, with the exception of the chosen estimation model. Here, I decided on the Generalized Partial Credit Model (GPCM). The GPCM presents 2 key parameters, discrimination parameter (a) and threshold parameter (b), where this model allowed the discrimination parameter to vary across items (Luo, 2018; Muraki, 1992; Muraki & Muraki, 2016), as the main difference to the Graded Response Model. Furthermore, the model is especially suitable for multidimensional IRT estimations (Chalmers, 2016; Muraki, 1992; Muraki & Muraki, 2016). The developed item bank provides the pool of questions for the CAT algorithm to choose from. As starting criteria a value of 0

for the ability score (θ) is defined, as this is common practice in CAT when the true ability score is unknown (Chalmers, 2016). As the starting rule, the Kullback-Leibler criteria was used as it is suitable for unidimensional and multidimensional data (Chalmers, 2016).

In CAT the next item is selected based on the theta score that best informs all the underlying latents simultaneously (Chalmers, 2016; Erdem Kara, 2019), which is in line with the assumption that the underlying latents are equally important. I defined the 'D-rule' as the item selection rule, as the most suitable selection to simultaneously increase the maximal amount of information and provide the largest matrix determinant (Chalmers, 2016). As the scoring rule, I introduced Maximum A Posteriori (MAP). This Bayesian scoring method estimates the latent trait by maximizing the posterior distribution, which combines the likelihood of the observed responses with a prior distribution of the latent trait, providing more stable and accurate ability estimates (Chalmers, 2016). I defined two termination criteria. I defined the test length with a maximum of items (i.e., 43 items), and based on measurement precision with a standard error that is defined as < 0.3 (Chalmers, 2016) and applies to all the latents simultaneously. The testing would stop when one or both of these criteria are matched. The maximal number of items is justified by the maximal number of open response items, which allows a fair comparison between the data compositions. I compared the CAT results to a sequential test, introducing the standard error as termination criteria and no restriction to the number of items. During the sequential testing, the simulation presents a new item until this criterion is matched without using a selection algorithm.

Evaluation

I use different approaches to evaluate the model performance. First, I compare how well the model fits the data based on the best model fit. Here, I examine the model fit indices AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) to inform my decision. For both fit indices, a lower value indicates better model fit. The usage of these fit indices is a well-established evaluation step in psychological research to ensure that the defined model aligns with the structure presented by the data. I developed three different models based on the item properties depicted in the item bank containing three different data compositions (i.e., rating scale items, open-response items, and a combination of both). I, furthermore, compared confirmatory, exploratory, and a bifactor model. I examined different models to find a middle ground between explanation (i.e., theorized model representation) and prediction (i.e., data defined structure).

Second, I inspect which model shows the best theta estimation (i.e., best estimation based on the smallest amount of items combined with the smallest standard error) to identify

one winning model. To achieve this, I systematically compared the performance of the selected adaptive models. The adaptive models are the rating scale, open response, and mixed model, emerged through the CAT simulations, using an algorithm to determine the next best item. For the model comparison, I calculated the total number of items used during the test. I follow the rationale, that the total number of items, administered in each model, is indicative of the model's efficiency. I compute the sum of squared differences (SSD) between the theta estimates of the adaptive models across all constructs over the delivered items to measures how much the theta estimates differ between the models. This allows to compare the consistency of the latent trait measurement over the three models, which can be used as an accuracy measurement. A smaller discrepancy shows similar estimates for the same construct, while larger discrepancies indicate greater differences in estimating the latent trait. To support this accuracy evaluation, I calculated the median standard error over the nine factors within the models and compared the reduction over the number of items. Additionally, I calculated the interquartile range (i.e., 25th to 75th percentile) to capture the variability across the nine different factors. Using this approach allows me to compare the model performance and evaluate the level of uncertainty.

Third, I extract the factor scores from the favorable model and correlate them with the sum scores of the established rating scales. In this way, I am validating the model with higher correlations suggesting that the ALIRT model captures the same underlying constructs.

As a last performance evaluation, I used the person fit statistics to calculate the percentage of individuals with an acceptable fit. Here, the criteria defined by the previously mentioned standards of person fit statistics.

Results

Sample Description

The presented sample was recruited from the general population. I pre-screened the participants to create participant pools for each diagnosis of interest. Here, only individuals who are stated to be diagnosed by a professional are included. The participants are assigned to one group based on their primary diagnosis. However, I also assessed if the participants have comorbidities, and indeed the majority of the participants have more than one diagnosis. For an overview see Table 1.

In the sample, 186 individuals are not receiving any treatment and 329 individuals receive one or more forms of treatment (i.e., medication and/or psychotherapy).

Table 1

Number of Participants with a Single Diagnosis

Diagnosis	Nr. of Participants
Alcohol/ Substance Abuse	1 (97.83 % with comorbidities)
Attention Deficit (Hyperactivity) Disorder (ADHD/ADD)	3 (93.48 % with comorbidities)
Autism Spectrum	4 (90.91 % with comorbidities)
Bipolar Disorder	6 (87.50 % with comorbidities)
Depression	23 (47.73 % with comorbidities)
Eating Disorder	0 (100% with comorbidities)
General Anxiety Disorder	14 (69.57 % with comorbidities)
Post-traumatic Stress Disorder	1 (97.87 % with comorbidities)

Note. Number of participants with only one diagnosis and the percentage of participants with more than one diagnosis for each group.

Prediction Model & Accuracy Assessment

The prediction model provided graded representations for the open responses asking for descriptive words. The model successfully converted the word embeddings and assigned values from 1 to 4 to the provided open responses, based on the probabilistic segments (25%, 50%, 75%, and 100%).

I evaluated the diagnostic accuracy of the rating scales, open responses, and narratives. The results indicate that open responses and rating scales are overall similar in their accuracy. Comparing precision, recall, and f1-scores the results indicate rather poor

performance for all response types. The precision score measures the accuracy of positive predictions, with higher scores indicating less false positive cases. The recall rate shows the ability of the algorithm to detect all actual true cases (i.e., completeness), with lower rates indicating a higher likelihood to miss positive cases. The f1-score balances precision and recall, with higher scores indicating a better balance between the two. The individual diagnoses show differences between the rating scales and open responses in precision, recall, and f1-scores. For example, the rating scales for ADHD, ASD, BID, PTSD, and SAD showed higher precision and therefore fewer false positive cases compared to open responses. While for GAD, Control, ED, and MDD the open responses are less prone to false positive categorizations. For a comprehensive overview see Table 2 and Figure 1.

The narratives showed a very low accuracy with 0.20 for the Mental Health Narrative and 0.19 for the Traumatic Event Narrative. The majority of the individuals have been predicted as control. Due to these poor results, I disregarded the narratives for further analysis during the current model.

Table 2

Accuracy Assessment Rating Scales, Open Reponses, & Mental Health Narrative

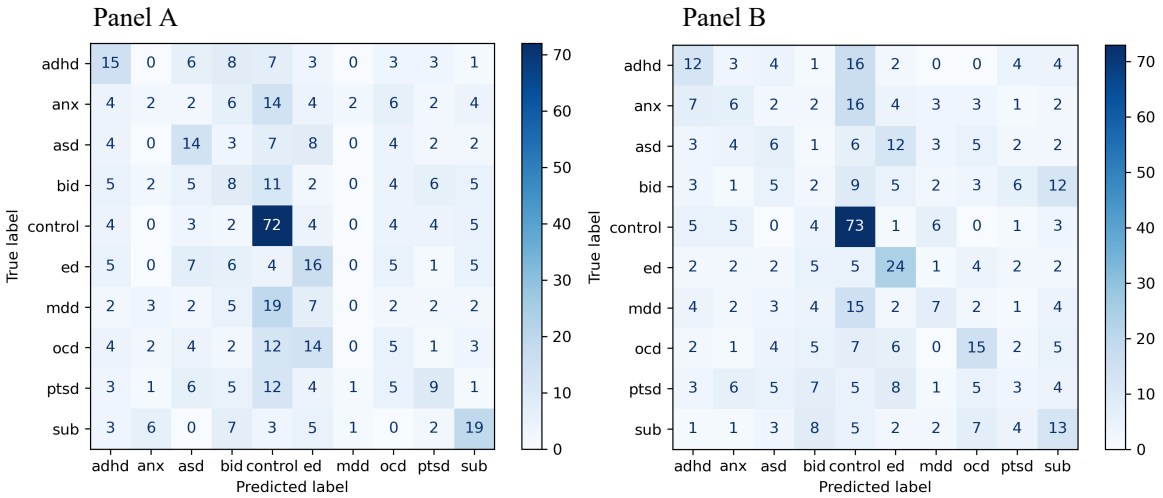
Scale	**Precision**			**Recall**			**f1-score**			**Support**
	RS	OR	N	RS	OR	N	RS	OR	N	
ADHD	0.31	0.29	0.00	0.33	0.26	0.00	0.32	0.28	0.00	46
GAD	0.12	0.19	0.00	0.04	0.13	0.00	0.06	0.15	0.00	46
ASD	0.29	0.18	0.00	0.32	0.14	0.00	0.30	0.15	0.00	44
BID	0.15	0.05	0.00	0.17	0.04	0.00	0.16	0.05	0.00	48
Control	0.45	0.46	0.21	0.73	0.74	0.99	0.56	0.57	0.35	98
ED	0.24	0.36	0.07	0.33	0.49	0.02	0.28	0.42	0.03	49
MDD	0.00	0.29	0.50	0.00	0.16	0.02	0.00	0.21	0.04	44
OCD	0.13	0.34	0.27	0.11	0.32	0.06	0.12	0.33	0.10	47
PTSD	0.28	0.11	0.00	0.19	0.06	0.00	0.23	0.08	0.00	47

SAD	0.40	0.25	0.33	0.41	0.28	0.02	0.41	0.27	0.04	46
	RS				OR				N	
Accuracy	**0.31**				**0.31**				**0.20**	

Note. Results showing precision, recall, f1 score, support for rating scales, open responses, and the mental health narrative respectively. These metrics are commonly used to evaluate the performance of classification models as the prediction model. Precision (i.e., proportion of true positive predictions among all instances predicted as positive) measures the model's ability to avoid labeling negative instances as positive. Recall (i.e., sensitivity or true positive rate) represents the proportion of true positive instances among all actual positive instances, reflecting the model's ability to identify all relevant cases. f1-score (i.e., harmonic mean of precision and recall) provides a single metric that balances both precision and recall. RS - rating scale. OR - open response. N - mental health narrative.

Figure 1

Confusion Matrices for Rating Scales & Open Responses



Note. Panel A - Rating Scales. Panel B - Open Responses. mdd - Major Depression Disorder. anx - Generalized Anxiety Disorder. bid - Bipolar Disorder. ocd - Obsessive Compulsive Disorder. adhd - Attention-Deficit/Hyperactivity Disorder. asd - Autism Spectrum Disorder. ed - Eating Disorder. sub - Substance Abuse Disorder. ptsd - Post-traumatic Stress Disorder.

Item Bank

For the item bank, I considered open response items based on descriptive words and rating scale items. The selection of the items are based on separate models for rating scales and open responses since the conducted EFA suggested separate factors for these manifest variables. The conducted unidimensional IRT models were performed to inspect and compare the established scales with the selected open response scales resulting from the developed open response items. I evaluated the model representation on different fit indices (see Table 3).

Table 3

Model Representation Selected Open Response Scales

Scale	CFI	TLI	RMSEA	SRMSR
ADHD	1.00	1.00	0.01	0.04
GAD	1.00	1.00	0.01	0.06
ASD	0.99	0.98	0.05	0.06
BID	–	–	–	–
ED	0.99	0.98	0.03	0.06
MDD	0.99	0.99	0.03	0.08
OCD	1.00	1.00	0.01	0.05
PTSD	1.00	1.00	0.01	0.06
SAD	0.99	0.99	0.04	0.08

Note. Table represents the overall model fit for the developed open response items. For the BID scale not applicable due to saturation of the model (i.e., too few degrees of freedom). CFI - Comparative Fit Index. TLI - Tucker Lewis Index. RMSEA - Root Mean Square Error of Approximation. SRMSR - Standardized Root Mean Square Residual. MDD - Major Depression Disorder. GAD - Generalized Anxiety Disorder. BID - Bipolar Disorder. OCD - Obsessive Compulsive Disorder. ADHD/ADD - Attention-Deficit/Hyperactivity Disorder. ASD - Autism Spectrum Disorder. ED - Eating Disorder. SAD - Substance Abuse Disorder. PTSD - Post-traumatic Stress Disorder.

The item bank resulted in a mix of open-ended questions and rating scale questions with a total number of 151 items showing acceptable item properties as previously defined. The evaluated items showed reasonable item properties, with discrimination values between 1 and 4. ICC and IIC suggested an acceptable representation of the ability level and the level of information. However, I am lacking good representations for the lower end of the scales. The graphical evaluations of the Test Information Curve and the Item Information Curve show a trend in all open response scales to discriminate better on the upper end of the latent (see Appendix Table A). I evaluated how the items load on the respective factors to ensure that the construct is well represented by the underlying manifest variables (i.e., items). Here, the item loadings indicate a relevant relationship with consistent lambda values (λ) ≥ 0.30. These results indicate that the assessed items are able to capture the intended construct meaningfully. I compared for each construct the latent scores with the simple number-correct-score to ensure the model is appropriately calibrated. Here, the results showed the

favored high correlations for all the constructs ranging between 0.97 - 0.98. For an overview of the newly developed items, please see Appendix Table A.

Item Response Modeling, Computerized Adaptive Testing, & Evaluation

To inform the CAT simulations, I explored the data within the IRT framework. I considered the informational value of rating scale items, open response items, and a composition of both. Here, I hypothesized, considering the item development based on the DSM-5 criteria, the data would map on 9 latent constructs reflecting the respective diagnoses. The conducted EFA for the mixed data composition is partially consistent with this theoretical assumption. The graphic evaluation of the screeplot is suggesting that a 7 to 9 factor solution is reasonable. The EFA for the open responses only suggests a single factor solution. In respect to the diagnostic criteria of the DSM-5, I remained with a 9-factor solution to reflect the latent construct. As defined in the first evaluation step, I compared the different models to assess if the data is presented well. I tested exploratory and confirmatory IRT's to identify the best data representation and found that for some models (e.g., confirmatory model), the number of participants are not able to sufficiently inform them. I was forced to disregard confirmatory IRT models for further analysis, since they showed negative degrees of freedom, indicating a lack of data points. Furthermore, I disregarded a bi-factor model that has been theorized as the most suitable representation of the data. This exploration of the data led to the conclusion that the three data compositions (i.e., rating scale items, open response items, and mixed items) are pragmatic representations to run CAT simulations and identify the best item compilation. Furthermore, it made a comparison of AIC and BIC redundant since for each data compilation, only a single model converged. These compositions are the simplest data representations possible in the multidimensional IRT framework. In these defined IRT models, the items are allowed to freely load on the latents. However, considering the ratio between individuals and items, this model is rather poorly informed and the fit indices indicate insufficient model information (i.e., insufficiently large sample to inform the model).

For every simulation, I specified the used items and the IRT model and compared this model with the sequential model. For a graphical overview of the theta estimation over the number of items for each simulation, see Figure 2.

I aimed to identify the model that provided the best theta estimation with the smallest number of items and the lowest standard error (SE) as criteria for the best model. The results show that the rating scale model shows the best theta estimation based on the smallest SE difference and therefore, is the favorable model considering accuracy (i.e., SE difference).

However, this result is accompanied by the fact that every model exhausted the maximum number of items (i.e., 43 items) and neither of the models are more efficient than the others.

I calculated the sum of squared differences for the theta estimation and thetas standard error to compare the model performance over the number of items (see Figure 3A). The results show that the rating scale model and the mixed model show a gradual increase indicating an adjustment to the estimates without showing a dramatic divergence showing that the latent trait estimates are similar for both models. They reach a plateau after 12 items, with a stable estimation until 30 items. This indicates that the most stable estimation lies within this item range. The open response model shows the lowest cumulative sum of squared differences (SSD) in theta estimates, indicating that it deviates less from the baseline theta values compared to the rating scale model and the mixed model. This shows that the open response model remains steady and updates the theta estimates less profoundly. The performance of the open response model shows that it does not efficiently reduce the standard error (SE), as shown in Figure 3B, following, an increased number of items leads to greater uncertainty. In Figure 3B, the rating scale model and the mixed model show a greater standard error reduction compared to the open response model. This slower reduction in SE reflects a weakness in refining the precision of the open response model and its estimates as more items are answered. Please see Figure 3A and 3B, for the accuracy, efficiency, and uncertainty measures over the number of items.

I explored how much informational value the open response items add to the assessment and reduced the termination criteria for the mixed model. The termination criterion was limited to a standard error < 0.4 and no limitation to the number of items was set. Here, the mixed model exceeded the entire item pool without reaching the defined criteria for all the latents, with factor 7 showing the highest standard error with SE = 0.69. For an overview of the final theta estimates, please see Table 4.

Table 4

Summary Theta Estimation and Standard Errors

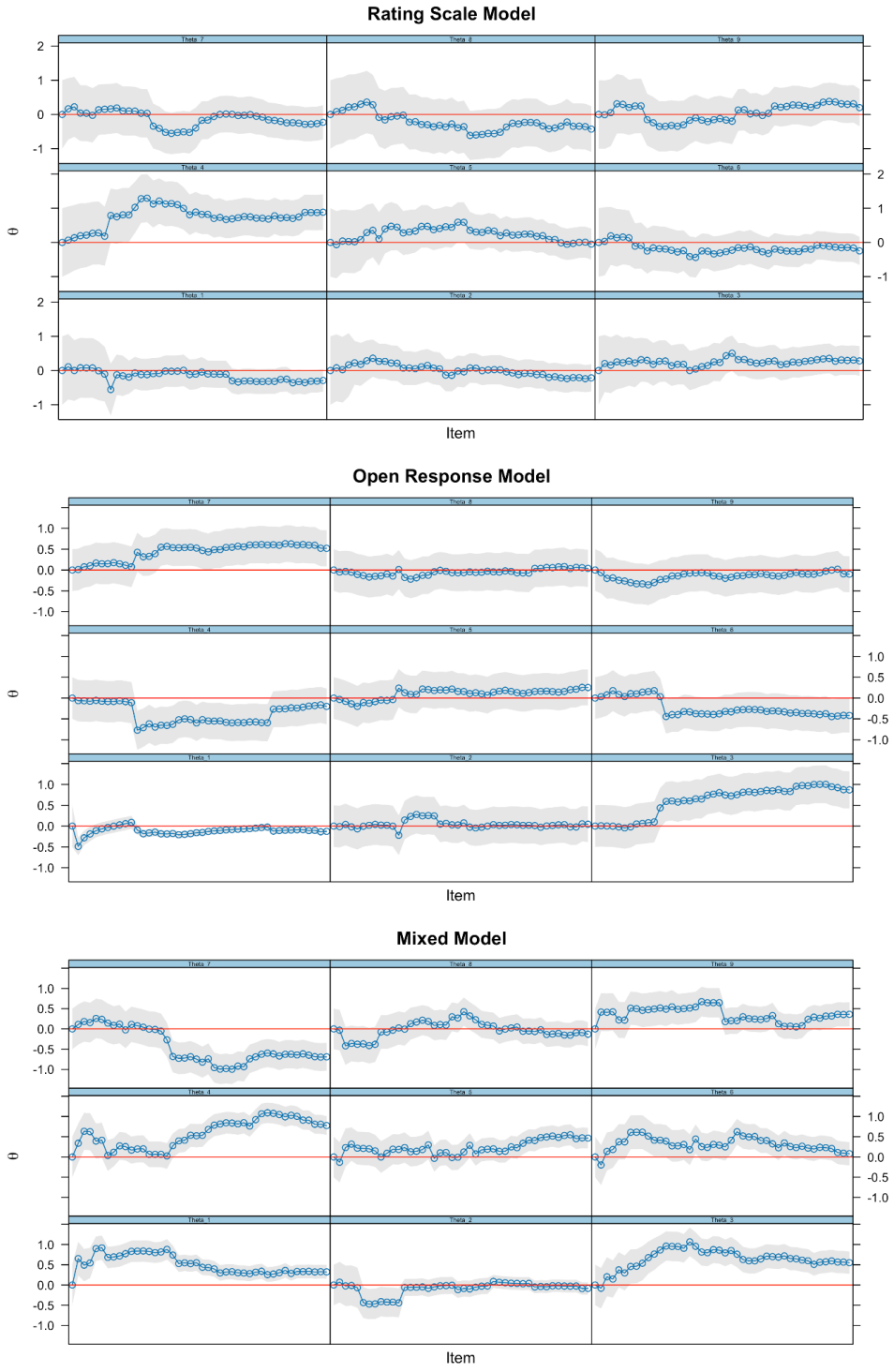
Model	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**
Mixed	0.33 (0.34)	-0.08 (0.30)	0.55 (0.55)	0.78 (0.46)	0.47 (0.51)	0.08 (0.57)	-0.69 (0.69)	-0.13 (0.60)	0.36 (0.59)
OR	-0.13 (0.14)	0.04 (0.85)	0.87 (0.89)	-0.20 (0.86)	0.25 (0.86)	-0.41 (0.82)	0.52 (0.87)	0.04 (0.89)	-0.10 (0.86)

RS	-0.29 (0.33)	-0.21 (0.40)	0.28 (0.43)	0.89 (0.52)	-0.05 (0.45)	-0.25 (0.41)	-0.23 (0.50)	-0.43 (0.67)	0.20 (0.54)
*Mixed**	0.16 (0.27)	0.08 (0.14)	-0.03 (0.32)	0.93 (0.35)	0.31 (0.38)	0.94 (0.43)	0.36 (0.46)	0.26 (0.45)	0.12 (0.49)

Note. Mixed - model including rating scale items and open response items. OR - model including open response items. RS - model including rating scale items. ‘’ marks the Mixed Model with the SE criterion as a single termination criterion.

Figure 2

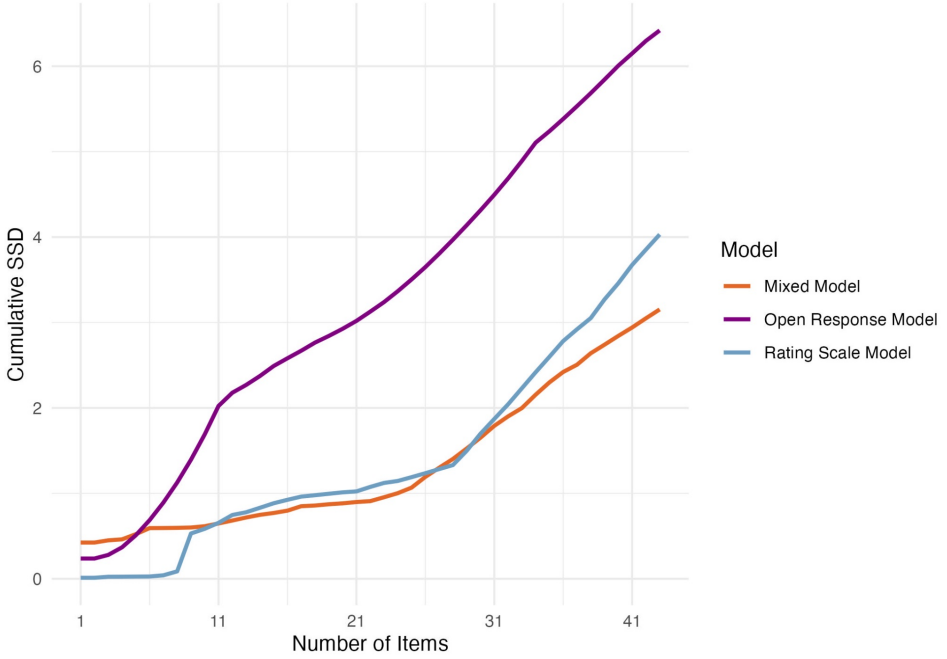
Theta Estimation over Number of Items



Note. Figure illustrates the performance of computerized adaptive testing simulations, comparing three types of item formats: rating scale items, open response items, and a combination of both within one model. The y-axis represents the standardized ability level (θ), the x-axis indicates the number of items, with a total of 43 items for each model. Each rectangle corresponds to a specific θ (i.e., a latent trait of interest), and each blue dot denotes an individual item. The shaded area reflects the 95% confidence interval. In each model, the next item is selected based on its informational contribution to the simultaneous estimation of θ.

Figure 3A

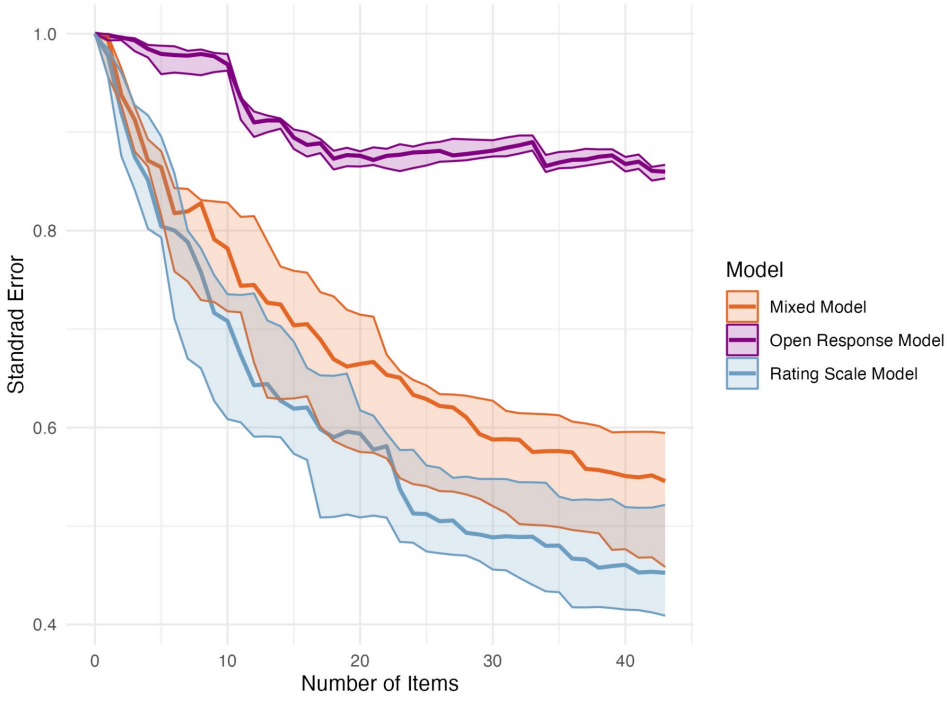
Accuracy as a Function of Number of Items



Note. Panel A displays the cumulative sum of squared differences (SSD) between the theta estimates of the three adaptive models (Rating Scale Model, Open Response Model, and Mixed Model) across the number of items. A greater divergence between the models indicates greater differences in estimating the latent trait.

Figure 3B

Efficiency Performance over Number of Items



Note. The figure illustrates the efficiency performance (i.e., the reduction of the standard error with every new item delivered) of each of the three models over the number of items. Efficiency is instrumentalized by the

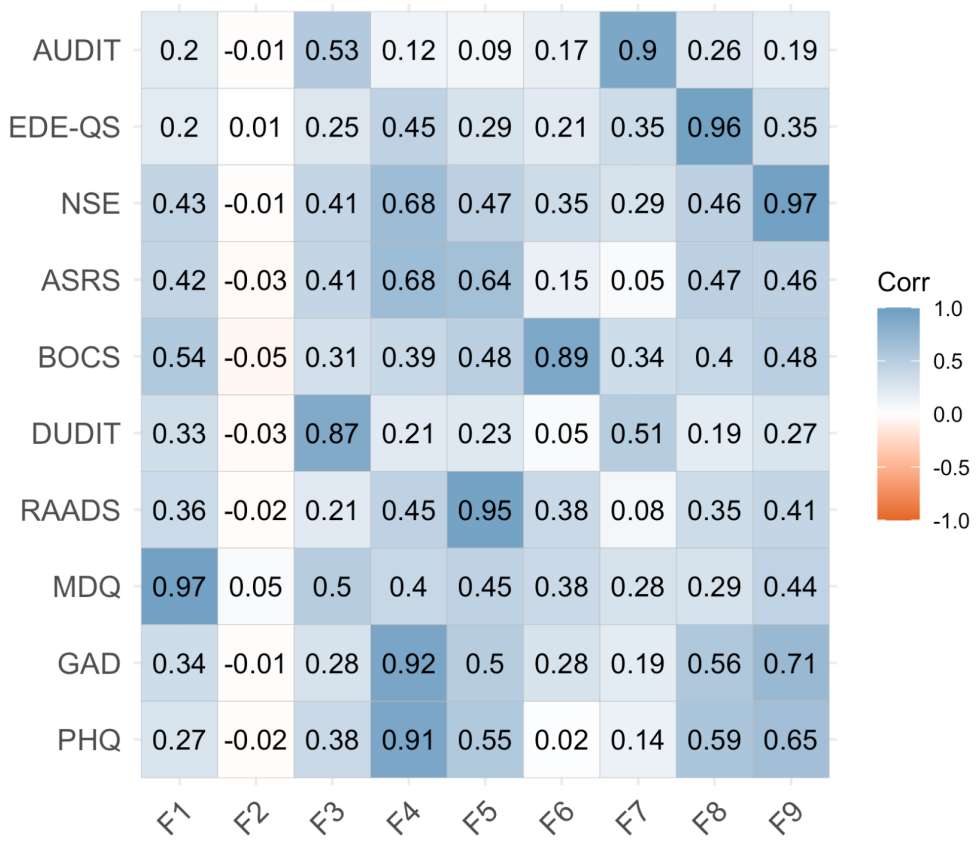
number of items. The standard error reduction supports the accuracy measurement showing the level of uncertainty within the estimates. The theta estimates standard errors are summarized for the nine factors, with the dark lines depicting the median standard error and the shaded regions representing the interquartile range (25th to 75th percentile) to capture the variability across the nine different factors.

I evaluated the models validity and aimed to gain a better understanding of the presented latent structure. I compared the estimated factor scores to the sum scores of the rating scales, where the factor scores present the ability estimation provided by the model. The results indicate that the established rating scales are partially reflected by the latent structure proposed by the EFA, since it shows high correlations between the factor scores and the rating scales (see Figure 4). However, since I was not able to sufficiently inform a confirmatory model, the presented structure differs from the theoretical informed structure (i.e., diagnostic criteria DSM-5). For one factor I don't see any relationship with the established measurements. This factor represents the open response items fully and does not correlate with the established measurements.

In the final evaluation step, I quantified the percentage of individuals with an acceptable fit. For the mixed model, the person fit indicates that ~ 14% of the individuals are well represented by the model. For the rating scale model ~ 53% of the individuals are well represented, and for the open response model, ~ 47% are well presented. This result supports the finding that the rating scale model is able to represent the participants' diagnosis most meaningfully.

Figure 4

Heatmap Factor Representations



Note. The correlation matrix shows the relationship between the established sum scores and the obtained theta estimations from the mixed model.

Participants Preferences

I assessed which response format is favored by the participants. Here, 109 participants preferred an open response format (35 participants preferred descriptive words and 74 writing a paragraph), while 220 participants preferred rating scales. 68 participants stated to not have a preference.

Discussion

The development of the ALIRT model provided important insights to further improve diagnostic practices and guide this process in the future. However, the presented results are discussed under certain limitations, following the conclusions are preliminary and will mainly provide guidance to improve the project and model development.

The developed open response items and the developed item bank provide a clinically suitable set of questions based on the DSM-5 criteria. Through the active engagement of clinical psychologists within the research group, I assured that the open response items and selected rating scales are meaningful from a clinical perspective. The diagnostic precision indicates that individuals are able to understand and respond to the developed items meaningfully. The quality of the rating scale items was confirmed. Focusing on the individual open response items and their response pattern, they consistently distinguish better on the upper end of the scale. Following, more pronounced symptoms are easier captured, while milder symptoms are not as well represented. I did not include any well-being or quality of life items, which might explain this underrepresentation. I suggest developing open response items focusing on positive experiences and protective factors, such as well-being, quality of life, and social support. In the future, this shortcoming needs to be addressed to cover the whole continuum of diagnostic categories and move further towards the aim of capturing the full experience of every individual.

Furthermore, while theoretically reasonable, the developed open response items do not provide a clear distinction between the diagnostic categories. I considered an exploratory and confirmatory approach, both showing that the newly developed items do not map on more than one factor. Analyzing the developed open response items in conjunction with the established rating scale items, both item groups load on different latens even though they theoretically present the same construct. One possible explanation could be that the observed patterns are due to the semantic overlap of symptom descriptions and the individual experience in general. Inconsistent with this explanation is that previous research showed that even disorders with similar symptoms (e.g., anxiety and depression) semantically differ (K. Kjell et al., 2021; O. N. E. Kjell et al., 2020; Sikström, Kelmendi, et al., 2023) and I would expect the same pattern here as well. The findings could be explained by considering that certain symptoms shape the individual experience extensively, masking or suppressing other symptoms. The sample specifications support this idea, showing that the majority of the

individuals have more than one diagnosis. This means that the sample shows such a great amount of comorbidities that the symptoms might not be clearly distinguishable by disorder with the developed items. In this case, the driving factor would not be the subjective experience of individual symptoms, but the actual presence of another disorder. Considering the high number of participants with comorbidities (ranges between 48% to 100% depending on diagnosis; see Table 1), it seems likely that this plays an important role. This finding would be an argument to move towards transdiagnostic mental health approaches, focusing on symptom clusters instead of diagnostic categories. Another potential covariate is that the majority of the individuals is receiving treatment (i.e., psychotherapy and/or medication), which influences the symptoms leading to potentially profound differences. This makes a clear representation even more difficult. In the future, expanding the sample to newly diagnosed, untreated individuals seems meaningful to disentangle the effect of treatments from the individual experience, or following participants over time to understand the development of the experience and the descriptive alterations.

Another explanation is the possibility that the prediction model might not provide a suitable grading. In this case, the misrepresentation would lie in the data modeling and not in the response pattern. The used prediction model reduces the high dimensional word embeddings to a graded scale by dividing the informational value of each word response into probabilistic segments. Even though polytomization is a pragmatic step to provide a data format that is simple to handle and has previously been successful (Varadarajan et al., 2023), it comes with costs. In this step, I trade the high dimensionality against simplification which automatically means a reduction of information. Moreover, I risk reintroducing the very same pitfalls that were initially criticized. For instance, in Item Response Theory (IRT), as with other multi-latent trait approaches, a common critique is the treatment of categorical scales as continuous (e.g., Kline, 2015). During the polytomization step, I convert a continuous into a categorical representation. Later I utilize packages that treat the categorical representations as continuous. This creates a circular problem, which may be more effectively addressed through an alternative approach in the first place. I suggest reflecting on what quality criteria are applicable to NLP-processed open responses. With the chosen approach, the high dimensional and informative open responses are compressed in a framework that might simply not be suitable to fairly present them. An improvement of the machine learning approach to overcome polytomization and find a more suitable way to use the data in an IRT framework is needed. A possible solution could be to use an optimization algorithm that is suitable for high dimensional data and allows the use of word embeddings without dimension

reduction. For example, the Particle Swarm Optimization has been suggested as a suitable algorithm for CAT (Zhehan, 2020) and is able to handle the high dimensionality of word embeddings. An integration of this technique might provide a solution to this limitation.

The misrepresentation could also lie within the language quantification itself. I used a BERT model to quantify the language, which creates contextual word embeddings. However, the provided descriptive words might not necessarily be contextual. The reasoning to represent the words with this model follows the assumption that *thinking* is contextual and embedded in the reflection of the subjective experience. This is an assumption that might not hold and which needs to be tested thoroughly. In the future, I suggest exploring which NLP approach provides the best representation of the provided responses. I suggest exploring models that are particularly trained to process language related to mental health as for example 'MentalBERT'. Potentially, decontextualized word embeddings such as word2vec or GloVe might be more suitable. These explorations might help to improve the language processing and present more reliable word embeddings.

I systematically evaluated and compared the performance of three adaptive models (i.e., rating scale model, open response model, and mixed model). This is instrumentalized through accuracy, efficiency, and uncertainty measures. Accuracy is instrumentalized as the sum of squared differences, with greater differences indicating greater differences in estimating the underlying latent traits. The results show, that the rating scale model and the mixed model show a small divergence indicating similar estimates. While the open response model clearly shows a greater difference, and in combination with the greater standard errors indicating a higher level of uncertainty. For the open response model, this means that with an increased amount of information the more uncertain the estimation gets. This seems reasonable on a theoretical level, assuming that individuals with mental health problems might become more descriptive when focused on less severe symptoms when provided with the option to extensively report about their experience. Considering the comorbidities in the sample, the 'fussy' representation of the latent might be a realistic representation of an subjective experience. This supports the idea, that open responses might capture symptom naunces that are not well represented in the described constructs. I explored my hypothesis that the ALIRT model would need fewer questions to assess mental disorders when compared to SRS. The results indicate that the rating scale model provides overall the most favorable representation, exhibiting the highest accuracy. I measured accuracy by testing which model achieved a sufficiently small standard error in estimating ability (theta) scores. Here, the rating scale model shows the smallest SE (see Figure 3B). These results are supported by the

person fit evaluation, indicating that the rating scale model aligns best with the expected response pattern. Interestingly, the smaller confidence intervals over the theta estimation on a factor level in the mixed model (Figure 2) signal a good performance of the open response model (see Figure 2). Furthermore, despite its higher accuracy, the rating scale model does not show a greater efficiency, which I instrumentalized by evaluating the number of items used in comparison to the other models. Every model needed the same amount of items during the CAT simulation. Following, I rejected the hypothesis that the ALIRT model would need fewer questions to assess mental health disorders compared to SRS. The model exceeded the entire item pool, without reaching a sufficiently small standard error. This finding could be explained in two different ways, that I suggest exploring in the future. First, the insufficient theta estimation might be caused by the model which does not well represent the sample structure as discussed previously. This would be in line with the fact that I had to disregard the theoretically most meaningful models (i.e., confirmatory model or bi-factor model). The chosen representation is a compromise between the mentioned limitations resulting in the most meaningful way to represent the data in light of the theoretical framework. In an exploratory multidimensional IRT, the manifest variables (i.e., items) are allowed to load freely on the factor with which they have the highest relationship. The greatest limitation of this model is that I was not able to specifically assign the items to a factor leading to unclear representation of the diagnoses. For example, the EFA indicates that depression and anxiety load on the same factor. According to DSM-5 criteria these are two distinguished diagnoses, even when symptoms overlap. This is a major limitation and shows that the current results can only give a first direction, but not provide valid answers yet. The exploratory multidimensional IRT model is not a suitable solution to accurately reflect mental disorders that are clearly defined and need to be respected in a diagnostic context.

In this current approach, I utilized diagnostic criteria of the DSM-5 as ground truth for the assessment and to describe the mental disorder of interest. These diagnostic categories are a simplification in itself and might not be the most accurate representation for the experience an individual with, for example, depression has. I suggest stepping further away from the diagnostic criteria and further exploring narrative responses. This is supported by partially convincing results of the precision analysis of the mental health narratives. For example for MDD, the mental health narrative showed a precision of 0.50, indicating that 50% of the responses categorized as positive were correctly predicted. This outperforms both, rating scales (precision score = 0.00) and open response items (precision score = 0.29). Until now these responses were not considered due to their overall limited diagnostic accuracy.

However, narratives might be most informative for the actual experience and should be explored in more detail. Recent research has shown that the while rating scale responses are more predictive for diagnoses, narratives are more predictive for actual behavior (Gu et al., n.d.). Considering these findings another rather traditional question in psychological research becomes relevant again- *What is the aim of the psychological assessment that is conducted?* I suggest distinguishing clearly between assigning a diagnosis or exploring the individual's experience and predicting future behavior. Depending on the question, one should consider setting a different focus. While a categorized psychological assessment is a pragmatic necessity in most European health care systems to, for example, get costs for mental health care reimbursed by health insurance. A defined diagnosis does not necessarily help an individual or psychotherapist in dealing with the related experiences. In this context, it might be more important to predict future behavior and personalize support and individualize treatments based on the quantified experience. In relation to this project, that means that I suggest extending the scope and exploring the potential of narratives, and if these are able to predict relevant behavioral patterns in a psychotherapeutic context.

I tested whether the combination of open-ended questions with traditional rating scales improved the validity of mental disorder categorization (i.e., SRS vs. OEQ) compared to a single approach. Focusing on construct validity, the analyses show that open response items and established rating scales do not represent identical constructs, suggesting that open responses capture unique aspects of mental health that are not reflected in traditional measures. Despite initial concerns regarding the precision of diagnoses derived from open response items, criterion validity assessments indicate that these items perform comparably to established scales, occasionally surpassing them in predictive accuracy. For example, the accuracy evaluations of the rating scales indicate that the PHQ9 scale was not able to assess depressive participants successfully. Indeed, the precision was 0.00 for the rating scale, while 0.29 for the open responses, indicating that 29% of the responses categorized as positive were correct compared to 0% for the rating scales.

The CAT simulations demonstrate that a mixed model, incorporating both item types, provides a comparably precise theta estimation, indicating the potential supporting informational value provided by open responses. This point is underlined by the reduced confidence interval in the mixed data composition compared to the rating scales and the open responses.

Considering previous findings, I hypothesized that open-responses would be preferred over SRS. This hypothesis was rejected as well, as the majority of participants preferred rating scales to descriptive words or writing paragraphs.

In summary, the results and their implications should be interpreted with caution due to several practical challenges. The previously mentioned need for advanced algorithms to accurately analyze the open responses, and the risk of potential response biases introduced through BERT are two of them. To draw valid conclusions and to provide directions for or against practical implementation further investigation is needed. The most crucial step would be a larger sample, and following further validation.

Limitations & Future Perspective

While the study offers important insights into the integration of natural language processing into item response theory, and the development of the ALIRT model, I want to expand on several limitations.

The developed prediction model trains the open responses on a single defined diagnosis, which neglects the potential richness of information that open responses can provide. Due to their high dimensionality, open responses contain nuanced data that can be relevant to multiple diagnoses. By focusing on only one diagnosis, the model is simplified and enhances its fit for that specific condition. However, this approach represents a clear trade-off, as it sacrifices the complexity and potential insights that could be gained from analyzing the responses in the context of multiple diagnoses. This limitation underscores the need for further development of the prediction model to account for the multidimensional nature of open-ended responses. In future developments, I aim to develop models that can handle the multidimensional aspects of mental health assessments. Incorporating advanced techniques such as multi-task learning or introducing algorithms that would allow the prediction model to leverage the rich information present in open-ended responses more effectively.

Even though power analysis has been performed to determine a sufficient sample size for the model, I realized that the complexity of the mental health representations call for a larger sample. The ALIRT model is under informed due to the small sample size, preventing me from drawing reliable conclusions. Several model specifications, as confirmatory or bi-factor models, were neglected due to a lack of data points to sufficiently inform them. To enhance the robustness and generalizability of the findings, I will increase the sample size in the future and further develop the model in collaboration with the research group. In the future, I will have the possibility to explore the most meaningful data representation that offers a compromise between a purely data driven approach (i.e., as I established it now) and a theory driven approach (i.e., based on predefined diagnostic criteria). This is crucial to combine the informational power of the rating scales and defined diagnoses, and add more nuanced information obtained by open responses. The presented latent structure suggests that the open responses might present unique mental health aspects that are not well reflected in the chosen scale. In the future it seems meaningful to expand on the investigated scale and to explore if this pattern is consistent over several scales, or if this is an artifact caused by the selected scales. As previous research showed, different depression scales capture different

aspects of the diagnosis (Fried, 2017). Following the same idea, it might be that the presented items capture different aspects of the diagnoses when compared to the SRS.

Due to the small sample, I did not account for the comorbidities of the participants during modeling, which most likely influenced the results. Considering the comorbid conditions in future research could provide a clearer understanding of the presented patterns, the models accuracy, and provide ecological validity during model training. This will provide a more comprehensive assessment framework that mirrors the complexity of real-world mental health scenarios.

A key aspect of the study involves the quantification of word representations using natural language processing. It is crucial to discuss and explore which embeddings are utilized and whether their size and complexity are appropriate for the task at hand. Large language models (LLMs) can sometimes be overly complex and may require careful evaluation to ensure they are suitable for the specific application (Bender et al., 2021).

When these limitations are addressed, there are various covariates that become meaningful to focus on in the future. For example, previous research has shown that language is influenced by different factors such as gender, age, and educational level (e.g., Meier et al., 2024; Newman et al., 2008). Another important aspect is the age-associated change of subjective experiences when it comes to mental health (Hegeman et al., 2012). Incorporating these factors into the model will be meaningful and can further inform the diagnostic process. Additionally, other potential covariates like socioeconomic status, cultural background, and cognitive abilities could provide a more comprehensive understanding of the influences on language on mental health assessment. Considering these factors within the model will not only enhance its accuracy but also provide deeper insights that can provide more effective and personalized diagnoses.

Conclusion

Humans naturally describe their state using common language rather than rating their thoughts and experiences on a numerical scale. This fundamental difference highlights a significant limitation of standardized rating scales (SRS), which often fail to capture the full depth and nuance of a person's experiences. SRS can be rigid and reductive, providing a constrained view that might not accurately reflect the individual's mental state.

The ALIRT model aims to overcome these limitations by allowing individuals to respond in their own words, thus providing a more nuanced and detailed representation of their mental health. This approach utilizes natural language processing to interpret open-ended responses, capturing the complexity and richness that traditional scales miss. By doing so, the ALIRT model offers a more flexible and nuanced assessment of mental health, aligning more closely with how individuals naturally communicate their feelings and experiences.

However, while the ALIRT model shows promise, I face many limitations which I need to address in the future. The ALIRT model represents a significant step forward, but substantial improvements are needed to fully realize its potential. These improvements include developing models that account for the multidimensional nature of mental health while respecting the latent structure given by the DSM-5, increasing the sample size, and addressing comorbidities. Furthermore, refining the natural language processing techniques used to quantify open-ended responses is essential to ensure they are both effective and appropriate for the specific application.

Most of these necessary improvements are discussed in this thesis and will guide the future development of the ALIRT model. By addressing these limitations and continuing to refine the model, I aim to create a more robust and versatile model for mental health assessment.

Data & Code Availability

The data and code is available to the examiners and the evaluator in the respective GitHub repository (https://github.com/rebeccaboe/MSc-ALIRT), which will be made public for the examination period. The repository includes a comprehensive read.me to guide the replication. Additionally, the developed stimulus material/ questionnaires (i.e., Qualtrics questionnaire) are available in the same repository.

References

Aditya Shastry, K., & Sanjay, H. A. (2023). Artificial Intelligence Techniques for the effective diagnosis of Alzheimer's Disease: A Review. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-023-16928-z

Agbavor, F., & Liang, H. (2022). Predicting dementia from spontaneous speech using large language models. *PLOS Digital Health*, *1*(12), e0000168. https://doi.org/10.1371/journal.pdig.0000168

Ahmad, F., Abbasi, A., Li, J., Dobolyi, D. G., Netemeyer, R. G., Clifford, G. D., & Chen, H. (2020). A Deep Learning Architecture for Psychometric Natural Language Processing. *ACM Transactions on Information Systems*, *38*(1), 6:1-6:29. https://doi.org/10.1145/3365211

Bejerot, S., Edman, G., Anckarsäter, H., Berglund, G., Gillberg, C., Hofvander, B., Humble, M. B., Mörtberg, E., Råstam, M., Ståhlberg, O., & Frisén, L. (2014). The Brief Obsessive–Compulsive Scale (BOCS): A self-report scale for OCD and obsessive–compulsive related disorders. *Nordic Journal of Psychiatry*, *68*(8), 549–559. https://doi.org/10.3109/08039488.2014.884631

Beltrami, D., Gagliardi, G., Rossini Favretti, R., Ghidoni, E., Tamburini, F., & Calzà, L. (2018). Speech Analysis by Natural Language Processing Techniques: A Possible Tool for Very Early Detection of Cognitive Decline? *Frontiers in Aging Neuroscience*, *10*. https://www.frontiersin.org/articles/10.3389/fnagi.2018.00369

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Broderick, M. P., Di Liberto, G. M., Anderson, A. J., Rofes, A., & Lalor, E. C. (2021).

Dissociable electrophysiological measures of natural language processing reveal differences in speech comprehension strategy in healthy ageing. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-84597-9

Butryn, T., Bryant, L., Marchionni, C., & Sholevar, F. (2017). The shortage of psychiatrists and other mental health providers: Causes, current state, and potential solutions. *International Journal of Academic Medicine*, *3*(1), 5. https://doi.org/10.4103/IJAM.IJAM_49_17

Chalmers, R. P. (2016). Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications. *Journal of Statistical Software*, *71*, 1–38. https://doi.org/10.18637/jss.v071.i05

Chowdhary, K. R. (2020). Natural Language Processing. In K. R. Chowdhary (Ed.), *Fundamentals of Artificial Intelligence* (pp. 603–649). Springer India. https://doi.org/10.1007/978-81-322-3972-7_19

Dawson, D. A., Grant, B. F., Stinson, F. S., & Zhou, Y. (2005). Effectiveness of the Derived Alcohol Use Disorders Identification Test (AUDIT-C) in Screening for Alcohol Use Disorders and Risk Drinking in the US General Population. *Alcoholism: Clinical and Experimental Research*, *29*(5), 844–854. https://doi.org/10.1097/01.ALC.0000164374.32229.A2

Del Re, A. C., Flückiger, C., Horvath, A. O., & Wampold, B. E. (2021). Examining therapist effects in the alliance–outcome relationship: A multilevel meta-analysis. *Journal of Consulting and Clinical Psychology*, *89*(5), 371–378. https://doi.org/10.1037/ccp0000637

Diaz-Asper, C., Chandler, C., Turner, R. S., Reynolds, B., & Elvevåg, B. (2022). Increasing access to cognitive screening in the elderly: Applying natural language processing methods to speech collected over the telephone. *Cortex*, *156*, 26–38.

https://doi.org/10.1016/j.cortex.2022.08.005

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*(1), 5–18. https://doi.org/10.1007/s11136-007-9198-0

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Psychology Press. https://doi.org/10.4324/9781410605269

Erdem Kara, B. (2019). Computer Adaptive Testing Simulations in R. *International Journal of Assessment Tools in Education*, *6*, 44–56. https://doi.org/10.21449/ijate.621157

Eriksson, J. M., Andersen, L. M., & Bejerot, S. (2013). RAADS-14 Screen: Validity of a screening tool for autism spectrum disorder in an adult psychiatric population. *Molecular Autism*, *4*(1), 49. https://doi.org/10.1186/2040-2392-4-49

Fatima, A., Li, Y., Hills, T. T., & Stella, M. (2021). DASentimental: Detecting Depression, Anxiety, and Stress in Texts via Emotional Recall, Cognitive Networks, and Machine Learning. *Big Data and Cognitive Computing*, *5*(4), Article 4. https://doi.org/10.3390/bdcc5040077

Fraser, K. C., Lundholm Fors, K., & Kokkinakis, D. (2019). Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Computer Speech & Language*, *53*, 121–139. https://doi.org/10.1016/j.csl.2018.07.005

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191–197. https://doi.org/10.1016/j.jad.2016.10.019

Gagliardi, G., Kokkinakis, D., & Duñabeitia, J. A. (2021). Editorial: Digital Linguistic Biomarkers: Beyond Paper and Pencil Tests. *Frontiers in Psychology*, *12*. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.752238

Gideon, N., Hawkes, N., Mond, J., Saunders, R., Tchanturia, K., & Serpell, L. (2018).

Correction: Development and Psychometric Validation of the EDE-QS, a 12 Item Short Form of the Eating Disorder Examination Questionnaire (EDE-Q). *PLOS ONE*, *13*(11), e0207256. https://doi.org/10.1371/journal.pone.0207256

Glaz, A. L., Haralambous, Y., Kim-Dufor, D.-H., Lenca, P., Billot, R., Ryan, T. C., Marsh, J., DeVylder, J., Walter, M., Berrouiguet, S., & Lemey, C. (2021). Machine Learning and Natural Language Processing in Mental Health: Systematic Review. *Journal of Medical Internet Research*, *23*(5), e15708. https://doi.org/10.2196/15708

Goretzko, D., Pham, T. T. H., & Bühner, M. (2021). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. *Current Psychology*, *40*(7), 3510–3521. https://doi.org/10.1007/s12144-019-00300-2

Graham, S. A., Lee, E. E., Jeste, D. V., Van Patten, R., Twamley, E. W., Nebeker, C., Yamada, Y., Kim, H.-C., & Depp, C. A. (2020). Artificial intelligence approaches to predicting and detecting cognitive decline in older adults: A conceptual review. *Psychiatry Research*, *284*, 112732. https://doi.org/10.1016/j.psychres.2019.112732

Gu, Z., Kjell, K., Schwartz, H. A., & Kjell, O. (n.d.). *Natural Language Response Formats for Assessing Depression and Worry with Large Language Models: A Sequential Evaluation with Model Pre-registration*. Retrieved July 22, 2024, from https://osf.io/p67db/download

Harahan, M. (2010). A Critical Look at the Looming Long-Term-Care Workforce Crisis. *Generations*, *34*(4), 20–26.

Haroz, E. E., Kane, J. C., Nguyen, A. J., Bass, J. K., Murray, L. K., & Bolton, P. (2020). When less is more: Reducing redundancy in mental health and psychosocial instruments using Item Response Theory. *Global Mental Health*, *7*, e3. https://doi.org/10.1017/gmh.2019.30

Hegeman, J. M., Kok, R. M., Mast, R. C. van der, & Giltay, E. J. (2012). Phenomenology of

depression in older compared with younger adults: Meta-analysis. *The British Journal of Psychiatry*, *200*(4), 275–281. https://doi.org/10.1192/bjp.bp.111.095950

Hildebrand, M. (2015). The Psychometric Properties of the Drug Use Disorders Identification Test (DUDIT): A Review of Recent Research. *Journal of Substance Abuse Treatment*, *53*, 52–59. https://doi.org/10.1016/j.jsat.2015.01.008

Hirschfeld, R. M. A., Williams, J. B. W., Spitzer, R. L., Calabrese, J. R., Flynn, L., Keck, P. E., Lewis, L., McElroy, S. L., Post, R. M., Rapport, D. J., Russell, J. M., Sachs, G. S., & Zajecka, J. (2000). Development and Validation of a Screening Instrument for Bipolar Spectrum Disorder: The Mood Disorder Questionnaire. *American Journal of Psychiatry*, *157*(11), 1873–1875. https://doi.org/10.1176/appi.ajp.157.11.1873

Høglend, P. (2014). Exploration of the Patient-Therapist Relationship in Psychotherapy. *American Journal of Psychiatry*, *171*(10), 1056–1066. https://doi.org/10.1176/appi.ajp.2014.14010121

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Igarashi, T., & Nihei, M. (2022). Cognitive Assessment of Japanese Older Adults with Text Data Augmentation. *Healthcare*, *10*(10), Article 10. https://doi.org/10.3390/healthcare10102051

Johnson, S. U., Ulvenes, P. G., Øktedalen, T., & Hoffart, A. (2019). Psychometric Properties of the General Anxiety Disorder 7-Item (GAD-7) Scale in a Heterogeneous Psychiatric Sample. *Frontiers in Psychology*, *10*.

https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01713

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31–36. https://doi.org/10.1007/BF02291575

Keetharuth, A. D., Bjorner, J. B., Barkham, M., Browne, J., Croudace, T., & Brazier, J. (2021). An item response theory analysis of an item pool for the recovering quality of life (ReQoL) measure. *Quality of Life Research*, *30*(1), 267–276. https://doi.org/10.1007/s11136-020-02622-2

Kessler, R. C., Adler, L., Ames, M., Demler, O., Faraone, S., Hiripi, E., Howes, M. J., Jin, R., Secnik, K., Spencer, T., Ustun, T. B., & Walters, E. E. (2005). The World Health Organization adult ADHD self-report scale (ASRS): A short screening scale for use in the general population. *Psychological Medicine*, *35*(2), 245–256. https://doi.org/10.1017/S0033291704002892

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Kim, S. Y., & Lee, W.-C. (2023). Several Variations of Simple-Structure MIRT Equating. *Journal of Educational Measurement*, *60*(1), 76–105. https://doi.org/10.1111/jedm.12341

Kishimoto, T., Takamiya, A., Liang, K., Funaki, K., Fujita, T., Kitazawa, M., Yoshimura, M., Tazawa, Y., Horigome, T., Eguchi, Y., Kikuchi, T., Tomita, M., Bun, S., Murakami, J., Sumali, B., Warnita, T., Kishi, A., Yotsui, M., Toyoshiba, H., … Mimura, M. (2020). The project for objective measures using computational psychiatry technology (PROMPT): Rationale, design, and methodology. *Contemporary Clinical Trials Communications*, *19*, 100649. https://doi.org/10.1016/j.conctc.2020.100649

Kjell, K., Johnsson, P., & Sikström, S. (2021). Freely Generated Word Responses Analyzed

With Artificial Intelligence Predict Self-Reported Symptoms of Depression, Anxiety, and Worry. *Frontiers in Psychology*, *12*. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.602581

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

Kjell, O. N. E., Kjell, K., Garcia, D., & Sikström, S. (2020). Prediction and Semantic Trained Scales: Examining the Relationship Between Semantic Responses to Depression and Worry and the Corresponding Rating Scales. In S. Sikström & D. Garcia (Eds.), *Statistical Semantics: Methods and Applications* (pp. 73–86). Springer International Publishing. https://doi.org/10.1007/978-3-030-37250-7_5

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling, Fourth Edition*. Guilford Publications.

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.

Lai, J., Cella, D., Chang, C.-H., Bode, R. K., & Heinemann, A. W. (2003). Item banking to improve, shorten and computerize self-reported fatigue: An illustration of steps to create a core item bank from the FACIT-Fatigue Scale. *Quality of Life Research*, *12*(5), 485–501. https://doi.org/10.1023/A:1025014509626

Lang, J. W. B., & Tay, L. (2021). The Science and Practice of Item Response Theory in Organizations. *Annual Review of Organizational Psychology and Organizational Behavior*, *8*(Volume 8, 2021), 311–338. https://doi.org/10.1146/annurev-orgpsych-012420-061705

LeBeau, R., Mischel, E., Resnick, H., Kilpatrick, D., Friedman, M., & Craske, M. (2014).

Dimensional assessment of posttraumatic stress disorder in DSM-5. *Psychiatry Research*, *218*(1), 143–147. https://doi.org/10.1016/j.psychres.2014.03.032

Levitt, H. M. (2021). Introduction to the special section: Questioning established qualitative methods and assumptions. *Qualitative Psychology*, *8*(3), 359–364. https://doi.org/10.1037/qup0000222

Li, Y., Masitah, A., & Hills, T. T. (2020). The Emotional Recall Task: Juxtaposing recall and recognition-based affect scales. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(9), 1782–1794. https://doi.org/10.1037/xlm0000841

Luo, Y. (2018). *Parameter Recovery with Marginal Maximum Likelihood and Markov Chain Monte Carlo Estimation for the Generalized Partial Credit Model* (arXiv:1809.07359). arXiv. https://doi.org/10.48550/arXiv.1809.07359

Meier, T., Mehl, M. R., Martin, M., & Horn, A. B. (2024). When I am sixty-four… evaluating language markers of well-being in healthy aging narratives. *PLOS ONE*, *19*(4), e0302103. https://doi.org/10.1371/journal.pone.0302103

Metarugcheep, S., Punyabukkana, P., Wanvarie, D., Hemrungrojn, S., Chunharas, C., & Pratanwanich, P. N. (2022). Selecting the Most Important Features for Predicting Mild Cognitive Impairment from Thai Verbal Fluency Assessments. *Sensors*, *22*(15), Article 15. https://doi.org/10.3390/s22155813

Mittal, A., Dumka, L., & Mohan, L. (2023). A Comprehensive Review on the Use of Artificial Intelligence in Mental Health Care. *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1–5. https://doi.org/10.1109/ICCCNT56998.2023.10308255

Moret-Tatay, C., Iborra-Marmolejo, I., Jorques-Infante, M. J., Esteve-Rodrigo, J. V., Schwanke, C. H. A., & Irigaray, T. Q. (2021). Can Virtual Assistants Perform Cognitive Assessment in Older Adults? A Review. *Medicina*, *57*(12), Article 12.

https://doi.org/10.3390/medicina57121310

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. https://doi.org/10.1177/014662169201600206

Muraki, E., & Muraki, M. (2016). Generalized Partial Credit Model. In *Handbook of Item Response Theory*. Chapman and Hall/CRC.

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender Differences in Language Use: An Analysis of 14,000 Text Samples. *Discourse Processes*, *45*(3), 211–236. https://doi.org/10.1080/01638530802073712

Organization, W. H. (2022). *World mental health report: Transforming mental health for all* [Report]. World Health Organization. https://archive.hshsl.umaryland.edu/handle/10713/20295

Patel, S. R., Basaraba, C., Rose, S., Van Meter, P., Wall, M. M., & Simpson, H. B. (2022). The brief obsessive-compulsive scale: Development and validation of a self-report (BOCS-SR). *Journal of Obsessive-Compulsive and Related Disorders*, *33*, 100730. https://doi.org/10.1016/j.jocrd.2022.100730

Penfold, R. B., Carrell, D. S., Cronkite, D. J., Pabiniak, C., Dodd, T., Glass, A. M., Johnson, E., Thompson, E., Arrighi, H. M., & Stang, P. E. (2022). Development of a machine learning model to predict mild cognitive impairment using natural language processing in the absence of screening. *BMC Medical Informatics and Decision Making*, *22*(1), 129. https://doi.org/10.1186/s12911-022-01864-z

Prnjak, K., Mitchison, D., Griffiths, S., Mond, J., Gideon, N., Serpell, L., & Hay, P. (2020). Further development of the 12-item EDE-QS: Identifying a cut-off for screening purposes. *BMC Psychiatry*, *20*(1), 146. https://doi.org/10.1186/s12888-020-02565-5

Sikström, S., Höök, A. P., & Kjell, O. (2023). Precise language responses versus easy rating

scales—Comparing respondents' views with clinicians' belief of the respondent's views. *PLOS ONE*, *18*(2), e0267995. https://doi.org/10.1371/journal.pone.0267995

Sikström, S., Kelmendi, B., & Persson, N. (2023). Assessment of depression and anxiety in young and old with a question-based computational language approach. *Npj Mental Health Research*, *2*(1), Article 1. https://doi.org/10.1038/s44184-023-00032-z

Steel, Z., Marnane, C., Iranpour, C., Chey, T., Jackson, J. W., Patel, V., & Silove, D. (2014). The global prevalence of common mental disorders: A systematic review and meta-analysis 1980–2013. *International Journal of Epidemiology*, *43*(2), 476–493. https://doi.org/10.1093/ije/dyu038

Varadarajan, V., Sikström, S., Kjell, O. N. E., & Schwartz, H. A. (2023). *Adaptive Language-based Mental Health Assessment with Item-Response Theory* (arXiv:2311.06467). arXiv. https://doi.org/10.48550/arXiv.2311.06467

Vos, T., Allen, C., Arora, M., Barber, R. M., Bhutta, Z. A., Brown, A., Carter, A., Casey, D. C., Charlson, F. J., Chen, A. Z., Coggeshall, M., Cornaby, L., Dandona, L., Dicker, D. J., Dilegge, T., Erskine, H. E., Ferrari, A. J., Fitzmaurice, C., Fleming, T., … Murray, C. J. L. (2016). Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: A systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, *388*(10053), 1545–1602. https://doi.org/10.1016/S0140-6736(16)31678-6

Watkins, M. W. (2018). Exploratory Factor Analysis: A Guide to Best Practice. *Journal of Black Psychology*, *44*(3), 219–246. https://doi.org/10.1177/0095798418771807

Yeung, A., Iaboni, A., Rochon, E., Lavoie, M., Santiago, C., Yancheva, M., Novikova, J., Xu, M., Robin, J., Kaufman, L. D., & Mostafa, F. (2021). Correlating natural language processing and automated speech analysis with clinician assessment to quantify speech-language changes in mild cognitive impairment and Alzheimer's

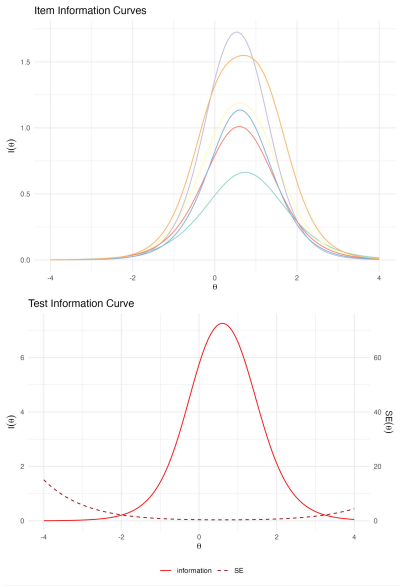
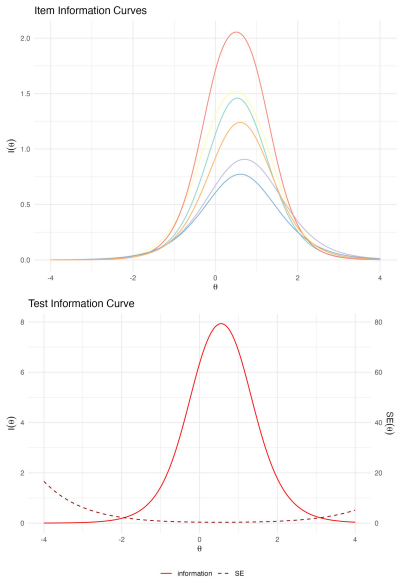
dementia. *Alzheimer's Research & Therapy*, *13*(1), 109. https://doi.org/10.1186/s13195-021-00848-x

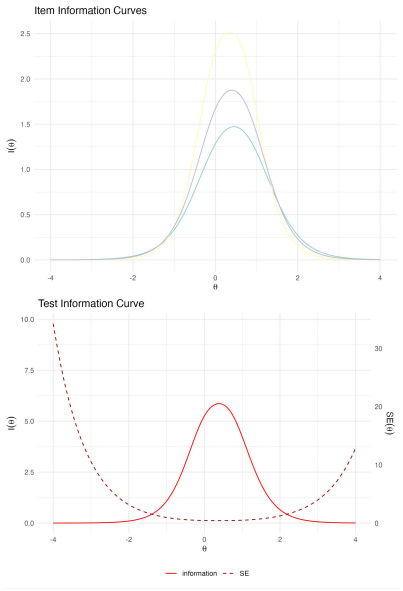
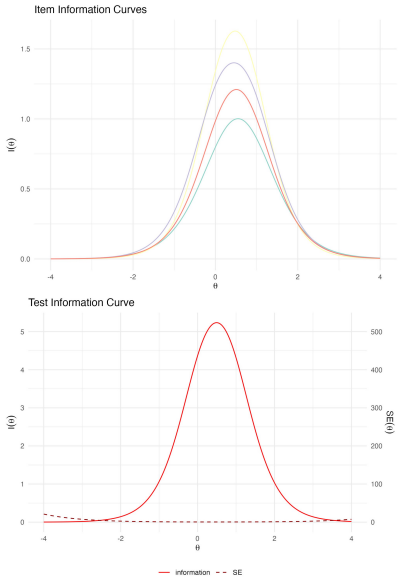
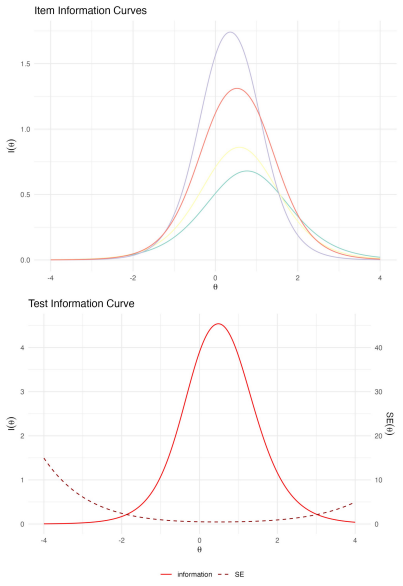
Zhehan, J. (2020). *Applying Particle Swarm Optimization To Estimate Psychometric Models With Categorical Responses*. https://kuscholarworks.ku.edu/entities/publication/5218fc90-afbe-4c72-b7dd-752c7e83b921

Appendix

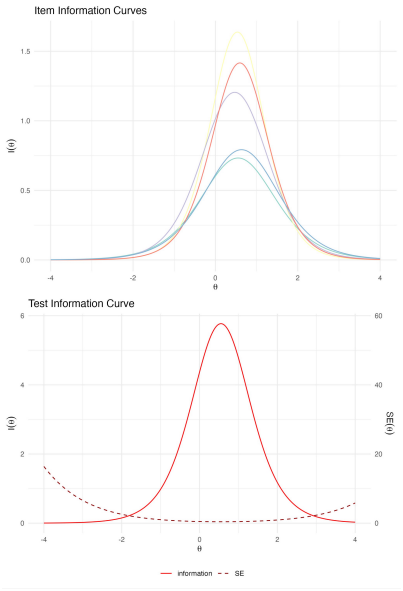
Table A

Item Bank with Open Response Items

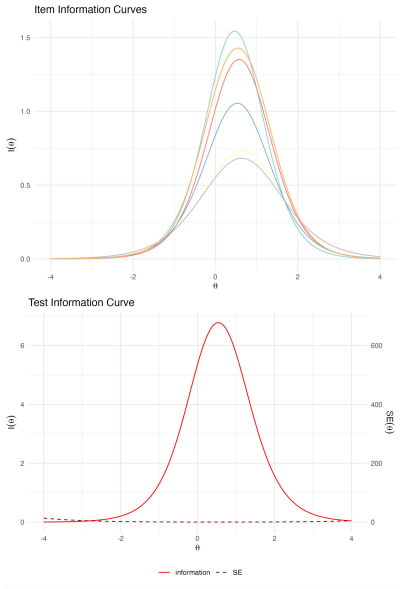
Item	Scale	a / λ	IIC & TIC
Describe **changes**, if any, in your **mood or emotions** in the past few weeks.	MDD	a = 1.56 λ = 0.68	
Describe a **persistent mood or emotions** you experienced in the past few weeks.	MDD	a = 1.96 λ = 0.76	
Describe your **ability to enjoy things** in the past few weeks.	MDD	a = 2.37 λ = 0.81	
Describe how your **appetite** has been lately.	MDD	a = 1.87 λ = 0.74	
Describe how your **sleep** has been lately.	MDD	a = 2.02 λ = 0.76	
Describe how your **motivation and/or energy level** has been lately.	MDD	a = 2.20 λ = 0.79	
Describe your **worries and their strength**, in the past few weeks.	GAD	a = 2.32 λ = 0.81	
Describe how your **mood has influenced your behavior** in the past few weeks.	GAD	a = 2.29 λ = 0.80	
Describe **places or activities you have avoided** due to anxiety.	GAD	a = 1.84 λ = 0.73	
Describe your **mental health during** the last two weeks.	GAD	a = 2.64 λ = 0.84	
Describe how your **mental health has influenced your behavior** in the past few weeks.	GAD	a = 1.76 λ = 0.72	
Describe **things** you have been **unable to do, concentrate on, make decisions on,** or carry out due to your mental health.	GAD	a = 2.15 λ = 0.78	
Describe **impulsive or risky behaviors** you have been engaged in lately.	BID	a = 2.22 λ = 0.86	
Describe how your **mood** has **influenced your daily life**, in the past few weeks.	BID	a = 2.92 λ = 0.83	

Consider your **main mental health symptoms**, **how long** have you been experiencing them?	BID	a = 2.49 λ = 0.79	
Describe **recurring thoughts** you experienced, and **their content**, in the past few weeks.	OCD	a = 2.08 λ = 0.78	
Describe **actions or rituals that you felt compelled to perform repeatedly**, in the past few weeks.	OCD	a = 2.21 λ = 0.79	
Describe **obsessive thoughts or compulsions** that you attempted to resist.	OCD	a = 2.38 λ = 0.81	
Describe how your **emotions and social relations** have been **influenced by your mental health**.	OCD	a = 1.93 λ = 0.75	
Describe your **attention during tasks or assignments**.	ADHD/ ADD	a = 2.05 λ = 0.77	
Describe **activities** of **restlessness, impulsivity,** and **ill-considered decisions**.	ADHD/ ADD	a = 2.39 λ = 0.82	
Describe how your **attention and activity level** have influenced your **social relationships**.	ADHD/ ADD	a = 1.73 λ = 0.71	
Describe how your **attention and activity level** have influenced your **work**.	ADHD/ ADD	a = 1.55 λ = 0.67	

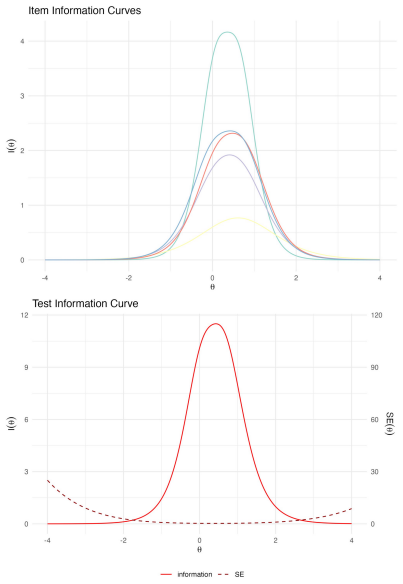
Describe your **typical social interaction**.	ASD	a = 1.68 λ = 0.70
Describe activities that you do to **avoid social interactions.**	ASD	a = 2.43 λ = 0.82
Describe **overwhelming** or **distressing sensory experiences**.	ASD	a = 2.05 λ = 0.77
Describe your **feelings** when faced with **changes and routine breaks**.	ASD	a = 2.28 λ = 0.80
Describe how you **experience social relationships**.	ASD	a = 1.71 λ = 0.71

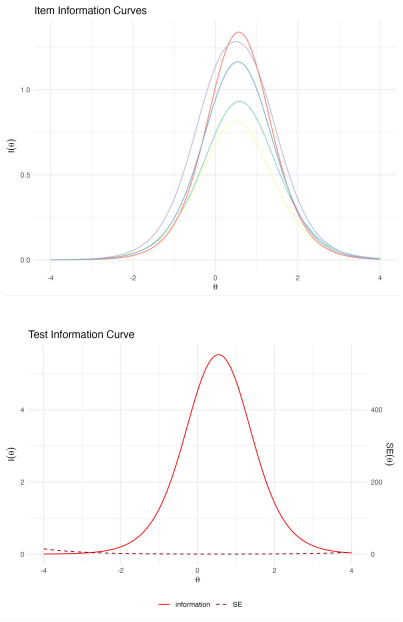


Describe your **eating habits** that **differ from other people**. Consider the last week.	ED	a = 2.29 λ = 0.80
Describe your **thoughts about food**.	ED	a = 1.61 λ = 0.69
Describe your **thoughts about your weight, shape, or appearance**.	ED	a = 1.54 λ = 0.67
Describe the **control over your eating behavior** and **related feelings**.	ED	a = 2.13 λ = 0.78
Describe **behaviors and emotions** you relate to **food**.	ED	a = 1.93 λ = 0.75
Describe the **impact** your **eating behaviors** have on your **daily life and relationships**.	ED	a = 2.17 λ = 0.79



Describe the **circumstances** under which you **use substances**.	SAD	a = 1.69 λ = 0.71
Describe your **thoughts, behavior, and feelings** when you are **not using** substances that you typically use.	SAD	a = 2.52 λ = 0.83
Describe **social, educational, or occupational consequences** you experienced due to your **usage** of **substances**.	SAD	a = 2.78 λ = 0.85
Describe **risky behavior** that you engage in during your **usage of substances**.	SAD	a = 3.84 λ = 0.91
Describe your **tolerance level** towards substances.	SAD	a = 2.79 λ = 0.85



Describe **impactful event(s)** you experienced and that are **still influencing your life**.	PTSD	a = 2.02 λ = 0.77	
Describe t**houghts, memories, or dreams** related to **impactful events** that are influencing your life.	PTSD	a = 2.17 λ = 0.79	
Describe how your **mental health has influenced your work performance** in the past few weeks.	PTSD	a = 1.73 λ = 0.72	
Describe how your **body felt** in the past few weeks. Think about **physical symptoms** that have relevance for you.	PTSD	a = 1.80 λ = 0.73	
When did you **first notice difficulties** in relation to your **main mental health issues**?	PTSD	a = 2.02 λ = 0.76	

Note. a - Item Discrimination. λ - Factor Loading. IIC - Item Information Curve. TIC - Test Information Curve. Rows marked with a gray background are excluded items. MDD - Major Depression Disorder. GAD - Generalized Anxiety Disorder. BiD - Bipolar Disorder. OCD - Obsessive Compulsive Disorder. ADHD/ADD - Attention-Deficit/Hyperactivity Disorder. ASD - Autism Spectrum Disorder. ED - Eating Disorder. SAD - Substance Abuse Disorder. PTSD - Post-traumatic Stress Disorder.