ALBA: Adaptive Language-Based Assessments for Mental Health

Vasudha Varadarajan[1], **Sverker Sikström**[2], **Oscar N.E. Kjell**[2], **H. Andrew Schwartz**[1]
[1]Department of Computer Science, Stony Brook University
[2]Department of Psychology, Lund University
{vvaradarajan,has}@cs.stonybrook.edu

Abstract

Mental health issues differ widely among individuals, with varied signs and symptoms. Recently, language-based assessments have shown promise in capturing this diversity, but they require a substantial sample of words per person for accuracy. This work introduces the task of Adaptive Language-Based Assessment (ALBA), which involves adaptively *ordering* questions while also *scoring* an individual's latent psychological trait using limited language responses to previous questions. To this end, we develop adaptive testing methods under two psychometric measurement theories: *Classical Test Theory* and *Item Response Theory*. We empirically evaluate ordering and scoring strategies, organizing into two new methods: a semi-supervised item response theory-based method (ALIRT) and a supervised *Actor-Critic* model. While we found both methods to improve over non-adaptive baselines, We found ALIRT to be the most accurate and scalable, achieving the highest accuracy with fewer questions (e.g., Pearson r ≈ 0.93 after only 3 questions as compared to typically needing at least 7 questions). In general, adaptive language-based assessments of depression and anxiety were able to utilize a smaller sample of language without compromising validity or large computational costs.

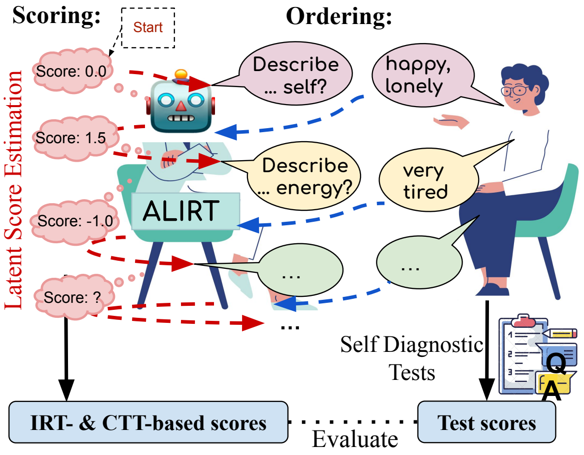

Figure 1: The ALBA task: the system picks the most informative question to ask based on previous responses, much like a therapist would in real life. To do this, we introduce an IRT-based semi-supervised method, ALIRT and an *Actor-Critic* model, and compare their performance with a limited set of language-response questions against self-report diagnostic questionnaire scores for depression and anxiety test scores (PHQ-9 and GAD-7).

1 Introduction

Standard mental health (e.g., depression, anxiety) assessment consists of asking patients a fixed set of questions to which they respond along a rating scale. For example, the Patient Health Questionnaire (PHQ) asks, "*Over the last 2 weeks, how often have you been bothered by having little interest or pleasure in doing things?* 0: Not at all, 1: Several days, 2: More than half the days, or 3: Nearly every day" (Kroenke et al., 2001; Siwek et al., 2009). This presents an information limitation: answering a fixed set of questions often leads to some unnecessary questions, while logging answers on a single dimensional rating scale limits the total information content possible.

Recent work has begun to address this information limitation by utilizing a patient's natural language to assess mental health conditions (Milne et al., 2016; De Choudhury et al., 2016; Eichstaedt et al., 2018; Kjell et al., 2019). Such open-ended language responses enable participants to elaborate on their mental health, where appropriate computational methods can be used to quantify the language response to a semantic scale (Kjell et al., 2019). Patients find that language is more precise in communicating their mental health issues, preferring it to rating scales (Sikström et al., 2023), but language-based assessments can be lengthy (Sikström et al., 2023) often requiring word minimums (Eichstaedt et al., 2021). Therefore, not only is there merit in mapping patients' language to their conditions but

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2466–2478
June 16-21, 2024 ©2024 Association for Computational Linguistics

also in asking questions or prompting for language more optimally to reduce redundant information and adapt iteratively to each question.

We introduce the task of Adaptive Language-Based Assessment (ALBA). ALBA is inspired by adaptive testing used in psychology and education (Xu et al., 2020), which usually utilizes a Bayesian statistical framework known as Item Response Theory (IRT) over discrete-valued responses to questions. It helps adapt to the future prompts based on the prompts already administered to the participant, iteratively picking the next best question to ask based on the current latent estimate.

We envision ALBA as a step towards the development of conversational diagnostic agents, as it allows for the modeling of language prompts in a semi-supervised manner. Using this approach, agents can adaptively conduct language-based assessments with dynamic scoring and benefit from a prior understanding of responses, leading to improved diagnostic accuracy, less use of patient and clinician time, and more personalized interactions.

Our main **contributions** include: (1) introducing the task of adaptive language-based assessment (ALBA) for selecting questions that provide the most informative responses; (2) development of ALIRT model, an approach integrating predictive modeling to be able to apply IRT to linguistic responses (as opposed to numeric responses as is typically used) – produces depression scores from only 4 question-responses that have 90% of the variance explained of an assessment that uses 11 question-responses; [1] (3) development of an *Actor-Critic* model model for adaptive language-based assessments; (4) evaluation of different modeling strategies within these models (e.g., discretizing in 2-tomous); (5) extensive empirical comparison of these methods with more straight-forward approaches covering multiple scoring strategies, and (6) insights into the questions that generally produce responses with the most information (i.e. ability to distinguish participant depression severity) informing better question/prompt creation.

2 Background

Item Response Theory estimates a latent variable as a proxy for an unobservable attribute (such as depression) by modeling the interaction between: (1) a latent variable (e.g. depression "score"), (2) observable variables from a population of participants (e.g. responses to questions [2] about lack of sleep, appetite, etc), and (3) the data points from a particular individual who is to be assessed. The latent score is typically estimated with Bayes parameter estimation. IRT addresses a short-coming of Classical Test Theory (CTT) (Lord and Novick, 2008) which is based on the assumption that there is a *true* score for an unobservable attribute, which is typically taken to be the summed value of all observable variables (e.g., ratings), and that the observed estimate is only off of the true score by some error from the act of measurement. This ignores correlations between the observable variables. In this work, we compare scores based on both IRT and CTT while evaluating the proposed adaptive testing methods over standard questionnaires for measuring major depression (PHQ-9) and generalized anxiety disorder (GAD-7).[3]

Various IRT modeling functions have been proposed for capturing discrete responses (Samejima, 2016; Muraki and Muraki, 2016; Chalmers, 2012). Increased model complexity in IRT leads to a rise in parameters, requiring larger datasets. Given the multidimensional nature of language representation and limited advances in simultaneous modeling across dimensions, we employ *polytomous* item response theory (Ostini and Nering, 2006) to discretize language responses onto a graded scale.

In this work, we introduce Adaptive Language-based IRT (ALIRT), which uses a supervised approach to polytomize the linguistic responses, and then employs adaptive testing using IRT.

3 Methods

The components that we develop for ALBA aims at (1) dynamically ordering a set of questions, picking next-best at a time; and (2) scoring the assessment at each step. These components can either be modeled jointly (as in ALIRT) or step-wise (as in our *Actor-Critic* model). We describe both approaches below and a suite of more straightforward baseline approaches to compare against.

[1] The code for ALIRT and other methods described in the paper can be found here.

[2] A *question* corresponds to an item in the Item Response Theory literature; therefore, the words "item" and "question" are used interchangeably in this paper.

[3] While the PHQ-9 and GAD-7 are still based on CTT, they are longer form questionnaires that check across symptoms described in the diagnostic and statistical manual, version 5 (American Psychiatric Association et al., 2013).

Algorithm 1 ALIRT

```
Notation: $J$: Num. of questions; K: Levels of rating scale;
    poly: polytomization split; tr: train split; te: test split;
    $M^j$: Model takes in item j embeddings, fit to "true" values;
    $D_i^j$: data split i; word embedding of response to item j.
    $Y_i$: Measure values (e.g. PHQ-9) for data split i.
    $\hat{Y}_i^j$: Predicted measures for data split i based on item j.
    $O_i^j$: data split i; polytomized response to item j.
    $\beta_j, \alpha$: Characteristic curve parameters for item j.

PHASE 1: POLYTOMIZATION

function POLYTOMIZATION(Train set: $\{D_{poly}, Y_{poly}\}$,
IRT train to polytomize:$D_{tr}$, IRT test to polytomize:$D_{te}$)
  for $j = 1, ..., J$ do
    $M^j \leftarrow$ FitRegression( $D_{poly}, Y_{poly}$)
    $\hat{Y}_{poly} \leftarrow M^j$.predict($D_{poly}$)
    $\hat{Y}_{tr}^j \leftarrow M^j$.predict($D_{tr}$) ; $\hat{Y}_{te}^j \leftarrow M^j$.predict($D_{te}$)
    thresholds $\leftarrow [(\frac{100k}{K})$-percentile$(\hat{Y}_{poly}); \forall k \in \{1,2...K\}]$
    $O_{tr}^j \leftarrow$ min(k) | thresholds[k] > $\hat{Y}_{tr}^j$ (likewise for $\hat{Y}_{te}^j$)
  end for
  return $O_{tr}, O_{te}$
end function

PHASE 2: ADAPTIVE TESTING

$\sigma_{IRT} = (\beta_1, \beta_2, ...\beta_J, \alpha) \leftarrow$ExpectationMaximization($O_{tr}$)
function GETNEXTQUESTION($\theta$, itemsLeft)
  return j | max. FisherInfo($\sigma_{IRT}^j(\theta)$) $\forall j \in$ {itemsLeft}
end function
function UPDATE(Item: j, Response: $o_{te}^j$)
  $\theta \leftarrow$ Maximum a Priori: maximize(P($o_{te}^j$ ,$\sigma_{IRT}$ | $\theta$) P($\theta$))
  return $\theta$
end function
for each set of user responses $o_{te}$ in $O_{te}$: do
  $\theta = \theta_0 \leftarrow 0$; itemsLeft $\leftarrow$ {1... J}
  while itemsLeft do
    j $\leftarrow$ GETNEXTQUESTION($\theta$, itemsLeft)
    $\theta \leftarrow$ UPDATE( j, $o_{te}^j$); itemsLeft $\leftarrow$ itemsLeft - {j};
  end while
end for
```

3.1 Adaptive Language-based IRT (ALIRT)

Adaptive Language-based IRT uses adaptive testing on language responses that are polytomized with a supervised model trained on word embeddings. The process is implemented in three phases:

Polytomization Language responses are multi-dimensional and can be represented as word embedding vectors. As discussed, we use *polytomous* item response theory to discretize the language responses to a graded scale. The responses are polytomized by training supervised models for each item on one split of the dataset: D_{poly}.

Word embeddings are extracted for each question and response in D_{poly}. For each question, the participants in our dataset (§4) are prompted for descriptive, context-independent words. Since contextual models aren't trained to represent individual descriptive words, we train our own word embeddings based on Principal Component Analysis over a term-document matrix with log-entropy weighting (aka Latent Semantic Analysis or LSA), allowing flexibility in choosing dimensions to represent language effectively. This approach has been proven as effective as word2vec, particularly in the context of psychology (Altszyler et al., 2016). The reduced dimensional space was 300, and the first 10 dimensions were used for the embeddings. The word embeddings were trained on a large dataset that contained similar word responses (69864 responses with 6728 unique words) to mental health questions– the dataset and word representations are introduced in (Kjell et al., 2019). Preferring smaller embedding sizes which are suitable for low-resource domains like mental health, we utilize 10 dimensions for each question. For replicability, any comparable word embeddings could be used: we explore dimension-reduction on GloVe and RoBERTA-large embeddings as well in Appendix A.

For each of the J questions, a multiple-ridge regression model is trained to predict the psychometric measure (PHQ-9 and GAD-7) on the averaged word embeddings of its responses. The average RMSE over all the ridge regression models is 10.93 for PHQ-9 and 8.64 for GAD-7. Each of the question's models is applied to the test set to predict the psychometric measures per question per sample. This is the *supervised* aspect of our approach.

The predicted psychometric measures on D_{poly} are thresholded based on percentiles given the discretization we wish to obtain for the questions. For example, if we want each question's responses to be polytomized from a scale of 1 to 3, i.e. [0,1,2,3], then the percentile thresholds are the quartiles for the predictions of the regression models for each question for D_{poly}. These thresholds are applied to the rest of the dataset, which is split into two more parts: D_{tr} for training the adaptive testing model, and D_{te} for evaluating the adaptive testing model.

Adaptive Testing with IRT can be directly applied to the polytomized data. In terms of Item Response Theory, an *item* is the question, and the corresponding *response* is the polytomized language response. To train the model, all the item parameters are simultaneously fit on D_{tr} using Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimization algorithm (Liu and Nocedal, 1989) as 2PL unidimensional IRT (Lord, 2012) until convergence (See Appendix D). We utilize a well-known R package for IRT, mirt[4]. For each data point in $D_{\texttt{te}}$, the testing is done by sequentially estimating the latent IRT variable for each question while keeping the learned IRT item parameters fixed. The latent variable is initialized at the average latent score of $D_{\texttt{tr}}$. To pick the next best question we utilize Fisher Information, a common criteria known to work well over most scenarios (Chalmers, 2016). The latent variable estimate ($\hat{L}$) is an *unsupervised* estimated value of the factor representing the selected questions.

To make the method comparable to other strategies based on the Classical Test Theory, we also calculate the average of the predicted measures ($\widehat{Y}$). For each question that is picked iteratively by the adaptive testing algorithm, to predict in the same scale as the psychometric measures.

We utilize mirtCAT[5], a computerized adaptive testing framework based on mirt to implement adaptive testing.

We run a 9-fold cross validation ($D_{\texttt{poly}}$:4, $D_{\texttt{tr}}$:4, $D_{\texttt{te}}$:1) across the two phases as described in the Algorithm 1– hence our approach is semi-supervised. Since the latent variable and the psychometric measures are on different scales, we report the Pearson r aggregated over all the nine test folds combined.

3.2 *Actor-Critic* model

Based on *Actor-Critic* framework used in the field of reinforcement learning (Grondman et al., 2012), we design a two-model system, where the first model (Measure Model) is guided by the second model (Error Model) to take the next step adaptively. Algorithm 2 provides a walk-through for this model. In our case, the Measure Model learns to predict the psychometric measures directly from the all the items administered so far, whereas the Error Model learns the error (MSE) of the Measure Model over each of the unadministered items. The Error Model dictates which item to select next based on the minimum predicted error. Unlike ALIRT, the prediction at each step does not depend on the previous step. The input to *Actor-critic* is predictions of the multiple ridge regression models for each question – a continuous value as opposed to the polytomized value in ALIRT.

We run a 9-fold cross validation with the same dataset split as ALIRT for comparability, such that $D_{\texttt{err}} = D_{\texttt{poly}}$, $D_{\texttt{tr}} = D_{\texttt{me}}$, and the test split $D_{\texttt{te}}$ being the same across experiments.

Algorithm 2 Actor-critic Adaptive Method

Notation: N: Number of folds, J: Number of questions
me: Measure split; err: Error split; te: Test split;
$M^{J'}_{\texttt{me}}$: Measure model– trained on responses to a set of items J', predicts the "true" score.
$E^{J'}_{\texttt{err}}[k]$: Error model– trained on responses to a set of items J', predicts error when k is the item to be added.
D^{j}_{i}: data split i; responses to item j; S: items administered
Y_i: Measure values (e.g. PHQ-9) for data split i.

```
TRAINING
function MEASUREMODELING($D_{me}$)
  for $j = 1, ..., J$ do
    $M^j \leftarrow$ FitRegression( $D_{me}$, $Y_{me}$)
    $\hat{Y}_{me} \leftarrow M^j$.predict($D_{me}$)
    $\hat{Y}^j_{tr} \leftarrow M^j$.predict($D_{tr}$) ; $\hat{Y}^j_{te} \leftarrow M^j$.predict($D_{te}$)
  end for
  for J'$\in$ powerSet(1,2 ... J) do
    $M^{J'}_{me} \leftarrow$ FitRegression( $D_{me}$, $Y_{me}$)
  end for
end function
function ERRORMODELING($D_{err}$)
  for $J_{temp} \in$ powerSet(1,2 ... J) do
    for ($i \in J - J_{temp}$) do # for each item not in $J_{temp}$
      $J' \leftarrow J_{temp}+\{i\}$
      $\hat{\delta}_{err} \leftarrow$ MeanSquaredError($Y_{me}$, $M^{J'}_{me}$.predict($D_{err}$))
      $E^{J_{temp}}_{err}$ [i] $\leftarrow$ FitRegression($D^{J_{temp}}_{err}$, $\hat{\delta}_{err}$ )
    end for
  end for
end function
ADAPTIVE TESTING
function GETNEXTQUESTION(itemsLeft, S)
  return j | min. $E^S_{err}$[j].predict($D^S_{err}$) $\forall j \in$ {itemsLeft}
end function
for each set of user responses $d_{te}$ in $D_{te}$: do
  $\theta = \theta_0 \leftarrow 0$; itemsLeft $\leftarrow$ {1... J}; S $\leftarrow$ {}
  while itemsLeft do
    j $\leftarrow$ GETNEXTQUESTION(itemsLeft, S)
    $\theta \leftarrow M^{S+\{j\}}_{me}$.predict( $d^j_{te}$ );
    itemsLeft $\leftarrow$ itemsLeft - {j}; S $\leftarrow$ S + {j};
  end while
end for
```

3.3 Baseline Models.

We experiment with different ordering strategies to compare the commonly used adaptive IRT criterion Maximum Fisher Information with traditional, baseline permutations of ordering. In particular, we explore three fixed-order approaches: **Random** – We use a random ordering of the questions for each participant, with any of the unasked questions having an equal probability of being asked next; **Forward Selection (fixedFor)** – As a fixed ordering baseline, we use forward selection to determine ordering, greedily picking the questions with the highest Pearson correlation for their polytomized item responses with the "true" scores; **Backward Elimination (fixedBack)** – As another fixed order-

[4] https://CRAN.R-project.org/package=mirt
[5] https://CRAN.R-project.org/package=mirtCAT

Model	Evaluated Against CTT					Evaluated Against $\hat{L}_{all}$					Num
	Num items					Num items					
	1	2	3	4	5	1	2	3	4	5	params
RandomOrder-$\hat{L}$	0.526	0.633	0.669	0.703	0.719	0.676	0.806	0.852	0.908	0.930	88
RandomOrder-$\widehat{Y}$	0.543	0.640	0.675	0.716	0.733	0.690	0.813	0.846	0.906	0.915	11
FixedBack-$\widehat{Y}$	0.600	0.669	0.701	0.738	0.749	0.787	0.877	0.908	0.932	0.945	11
FixedFor-$\widehat{Y}$	0.599	0.690	0.709	0.731	0.746	0.785	0.880	0.911	0.932	0.949	11
DecisionTree-$\hat{L}$ †	0.604	0.658	0.679	0.707	0.722	0.760	0.831	0.890	0.917	0.934	479
DecisionTree-$\widehat{Y}$	0.621	0.685	0.717	**0.740**	0.748	0.774	0.837	0.897	0.915	0.924	391
ActorCritic-$\hat{L}$†	0.585	0.656	0.668	0.699	0.719	0.748	0.828	0.881	0.908	0.929	11,341
ActorCritic-$\widehat{Y}$	0.619	**0.693**	0.714	0.739	**0.752**	0.765	0.841	0.893	0.914	0.926	11,253
ALIRT-$\hat{L}$	0.612	0.669	0.707	0.723	0.731	0.816	**0.897**	**0.935**	**0.955**	**0.965**	88
ALIRT-$\widehat{Y}$ †	**0.630**	0.685	**0.726**	0.739	0.746	**0.828**	0.895	0.929	0.949	0.955	88

Table 1: Performance at depression severity assessment across ordering and scoring strategies (Pearson r). We find adaptive testing to be better than fixed ordering, and considering parameter explosion, ALIRT is better. Methods suffixed by $\hat{L}$ utilize IRT for scoring (i.e. the latent variable), while those suffixed by $\widehat{Y}$ utilize a direct estimate for scoring ($\widehat{Y} = mean(\hat{y})$ for all y across administered questions). We find that the measures are consistent across both approaches. $\hat{L}_{all}$ refers to the latent score when all the 11 items are used. This means that administering just 3 items in the questionnaire based on ALIRT can achieve > 0.9 correlation (Pearson r) with the latent score from using all the 11 items in the questionnaire. †: Significant reduction in error ($p < .05$) across multiple tests, compared to the best baseline (FixedFor-$\widehat{Y}$). The p-values for all the correlations is < 0.001.

ing baseline, we use backward elimination based on eliminating items with the lowest Pearson correlations of responses with the “true” scores.

As a more sophisticated, adaptive baseline, we also explore the **Decision Tree**, which can be seen as defining an adaptive strategy where the next best question (“feature” in a decision tree) is picked based on the condition encountered at the current node. A decision tree is similar to IRT in that it can select the next best question contingent on previous responses. They are different in that the best splits are pre-calculated, and the next question is picked based on responses (“feature values”) at a node, whereas, for IRT, maximum Fisher information of the item parameters over all the remaining questions decides the next best question.

3.4 Scoring Paradigms

We also compare across two scoring paradigms across all the experiments (Tables 1, 2, 4). **Latent estimate ($\hat{L}$)** is the latent variable produced by the Item Response Theory (IRT) model. As the best latent estimate for depression (or anxiety), we consider the *most informative* latent estimate to derive from *all* the questions: $\hat{L}_{all}$) to evaluate the rest of the methods against. **Classical Test Score ($\widehat{Y}$)**, on the other hand, is the average of item scores, much like scores derived from a traditional questionnaire for mental health assessment, based on Classical Test Theory (CTT). In this work, we use the PHQ-9 (GAD-7) for depression (anxiety) severity as the CTT-based score to evaluate against.

Model	Outcome	1	2	4
RandOrder-$\widehat{Y}$	CTT	0.491	0.636	0.675
FixedFor-$\widehat{Y}$	CTT	0.598	0.638	0.694
DecTree-$\widehat{Y}$	CTT	0.581	0.643	0.664
ActorCritic-$\hat{L}$	CTT	0.561	0.631	0.672
ActorCritic-$\widehat{Y}$	CTT	0.587	0.658	0.705
ALIRT-$\hat{L}$	CTT	0.600	0.653	0.694
ALIRT-$\widehat{Y}$	CTT	**0.608**	**0.663**	**0.707**
RandOrder-$\widehat{Y}$	$\hat{L}_{all}$	0.603	0.770	0.902
FixedFor-$\widehat{Y}$	$\hat{L}_{all}$	0.805	0.877	0.935
DecTree-$\widehat{Y}$	$\hat{L}_{all}$	0.760	0.827	0.844
ActorCritic-$\hat{L}$	$\hat{L}_{all}$	0.740	0.841	0.906
ActorCritic-$\widehat{Y}$	$\hat{L}_{all}$	0.758	0.865	0.929
ALIRT-$\hat{L}$	$\hat{L}_{all}$	0.812	**0.904**	**0.958**
ALIRT-$\widehat{Y}$	$\hat{L}_{all}$	**0.818**	0.901	0.952

Table 2: Results for applying ALIRT for **anxiety severity**, as measured by GAD-7 as CTT- and $\hat{L}_{all}$-based scores. The reported values are Pearson r correlations with all the p-values < 0.001.

Across all the experiments described, the folds are kept consistent (including baselines).

4 Dataset

Our dataset consists of open-ended language answers to eleven questions and two self-diagnostic tests in the form of closed-ended rating scales. Participants were recruited online from Mechanical

Open-Ended Questions	Shorthand	Word-Response Correlation with the PHQ-9
Describe how you generally felt the last 2 weeks, that is, how you felt on average.	Describe Mental Health	0.61
Describe how you have been feeling about yourself over the last 2 weeks.	Describe Yourself	0.61
Over the last 2 weeks, have you been depressed or not?	Describe Depression or Not	0.58
Over the last 2 weeks, have you been worried or not?	Describe Worry or Not	0.41
Overall in your life, are you in harmony or not?	Describe Harmony or Not	0.54
Overall in your life, are you satisfied or not?	Describe Satisfaction or Not	0.57
Describe the nature of your physical movements over the last 2 weeks (have you for example been moving and speaking slowly; or the opposite, been fidgety and restless).	Describe Movement	0.45
Describe your sleep over the last 2 weeks.	Describe Sleep	0.44
Describe your concentration over the last 2 weeks.	Describe Concentration	0.44
Describe your appetite for food over the last 2 weeks.	Describe Appetite	0.36
Describe your energy level over the last 2 weeks.	Describe Energy	0.51

Table 3: The questions administered to participants in our dataset, along with their shorthand used in this paper. Pearson r is reported for each of the question's word response scores with the self-reported PHQ-9 scores.

Turk (N = 528; 2018-05-05) and Prolific (N = 419; 2018-11-28), where they were paid $3 and £3, respectively to participate. The MT data set included attention checks (e.g., "On this item, answer alternative 3"), which 64 participants failed and thus were removed. The Prolific study included a screening procedure where 260 participants had reported being diagnosed with Major Depressive Disorder and/or Generalised Anxiety Disorder before being invited to participate; 159 participants were not screened. The open-ended questions, shown in Table 3, concerned mental health and well-being, including: 1. General mental health, 2. Depression, 3. Anxiety, 4. Harmony, 5. Satisfaction; and some mental health-related symptoms: 6. Movement, 7. Sleep, 8. Concentration, 9. Appetite, 10. Energy, and 11. Self-perception. The participants were asked to respond using at least five descriptive words for the mental health questions (1-3), three descriptive words for the well-being questions (4-5), and two descriptive words for the symptom questions (6-11). The mean PHQ-9 score across the cohort was 11.98 with a standard deviation of 7.76, and the mean GAD-7 was 10.16 with a standard deviation of 6.23. The distribution of participants across PHQ-9 and GAD-7 scores are given in Figure 2. The validated scale for depression was the Patient Health Questionnaire 9-item aka the PHQ-9 (Kroenke et al., 2001), and for anxiety, the Generalised Anxiety Disorder 7-item scale aka the GAD-7 (Spitzer et al., 2006).

We model a single latent score in this dataset, as opposed to modeling multiple mental health conditions simultaneously with multidimensional IRT models (Chalmers, 2012). This choice is justified in Appendix C.

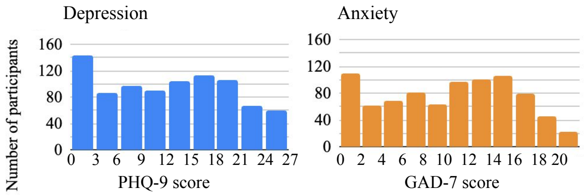

Figure 2: Distribution of depression and anxiety scores of participants in the dataset described in §4.

5 Results & Discussion

We report the performances of the various adaptive strategies, in comparison with the baselines and across scoring methods, in Table 1.

5.1 Adaptive Strategies

Table 1 examines the difference between two scoring methods with experiments run for depression severity assessment. For CTT-based scoring, we use a simple averaging of predicted measures over the selected questions, which is limited by the accuracies of the 11 individual question models ($\widehat{Y}$). $\hat{L}$ is the latent estimate produced by the IRT model. Each of these is evaluated against a "true" CTT score (PHQ-9) and a "true" latent score which is the latent estimate obtained by simultaneous parameter estimation with all the questions ($\hat{L}_{all}$). We find that adaptive strategies tend to perform bet-

Eval against:	CTT		$\hat{L}_{all}$		Num
Approach	**2**	**4**	**2**	**4**	**Params**
Actor-Critic					
$\widehat{Y}$	0.693	0.739	0.841	0.914	11,253
Regr ($\hat{Y}$)	**0.694**	**0.741**	0.873	0.927	11,253
Regr (X)	0.693	0.734	0.857	0.922	112,530
ALIRT					
$\hat{L}$	0.669	0.723	**0.897**	**0.955**	88
Regr ($\hat{Y}$)	0.685	0.736	0.896	0.947	11,341
Regr (X)	0.685	0.740	0.883	0.934	112,618

Table 4: Performance of depression severity assessment across ordering strategies for regression-based scoring strategies. We compare using regression on the item scores and on the word embeddings to the best from Table 1– $\widehat{Y}$ and $\hat{L}$. The reported values are Pearson r, with p-values < 0.001.

ter than the baselines. Among the three adaptive strategies used to directly predict the psychometric measures in Table 1, we find that the ALIRT-$\widehat{Y}$ performs best when compared against the CTT score, and ALIRT-$\hat{L}$ performs best when compared to $\hat{L}_{all}$. The differences in correlations become less evident as the number of items administered increases due to the convergence of items picked across different strategies.

The *Actor-Critic* model has $2^N - 1$ score prediction models and $N.(2^{N-1} - 1)$ error prediction models (see Appendix B) trained on each combination of questions to pick out the best question to administer. Despite this, the performance boost that could be afforded by the computational complexity is not always significant, and ALIRT performs similarly (or better) despite a much smaller number of parameters and shorter runtime. It is notable that ALIRT, which uses Maximum Fisher Information for adaptive ordering, does *not* try to optimize for errors/correlation with the "true" scores, but the ordering produced by it largely helps across both the scoring paradigms, which demonstrates the utility of IRT in being able to capture inherent associations without direct supervision.

These findings are fairly consistent for anxiety severity assessment as well, evaluated against GAD-7, as seen in Table 2. This indicates that adaptive language-based assessment could be extended to other common, standardized assessments as well.

5.2 Scoring strategies

We note from table 1 that there is merit to both the scoring paradigms, with CTT offering a widely accepted, standardized, fixed scale with supervision in every step, whereas IRT allows semi-supervision and can adapt the scale according to the response behavior of the cohort of participants. We compare the two scoring strategies to regression-based scoring as well, where instead of averaging the scores over the selected questions, we use regression to train prediction models to output a score, with the item response as input. Table 4 compares the various scoring strategies and how they correlate to CTT-based "true" scores and IRT-based "most informative" scores.

Regression over word embeddings – Regr(X) The input is the item response word embeddings. We find that this method does not really fare better across both *Actor-Critic* and ALIRT. Since we use 10 dimensional word embeddings, the number of parameters is increased tenfold, which could cause the model to overfit. Moreover, the method is unrealistic when scaled up to more questions due to parameter explosion.

Regression over predicted scores – Regr($\hat{Y}$) Item response scores are used as input to the model, and thus we can re-use all the models trained for *Actor-Critic* approach. While there is still risk of parameter explosion if there were more questions, the method does not demand more compute and seems to improve the correlations of the predicted scores in the *Actor-Critic* setting.

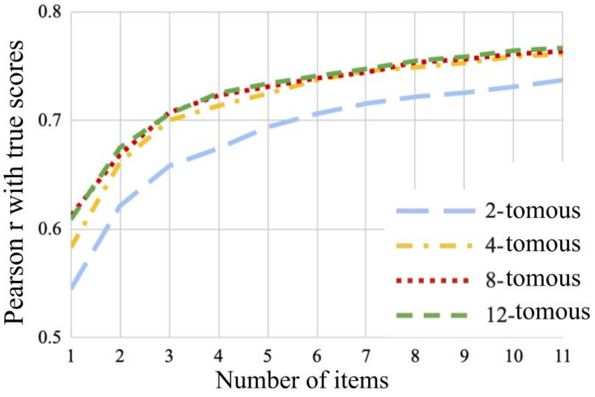

Figure 3: The correlation of the latent scores with the "true" (PHQ-9) scores for various polytomization levels across the number of items. 12-tomous model is likely to be overfit and does not offer significant advantage over our initial choice of 8.

5.3 Optimal Discretization

Some information can be lost when numeric values are polytomous (Catlett, 1991). For the purposes of use in adaptive testing with IRT, it is unclear how discretized the values should be. On the one hand, there can be more information loss with coarse-grained discretization (i.e. less number of choices in the rating scale); and on the other hand, fine-grained discretized (i.e too many choices in the rating scale) results in too many parameters with respect to data size.The result of our experiment is seen in Figure 3 where we experiment for 2, 4, 8 and 12. A polytomization of 8 works just as well as 12 with $\frac{2}{3}$ the number of parameters. We also found that a polytomization of 13 or above results in missing values – resulting in ill-fit characteristic curves used in the IRT model.

5.4 Most Informative Questions: Depression Severity

Based on Table 1, we find that ALIRT achieves a high correlation ($r > 0.7$) to standardized assessments with 3 questions. Figure 4 tells us that the questions are not highly personalized– for the first 4 items, only 6 out of total 11 questions are administered, with general mental health questions ("Describe Yourself" and "Describe Mental Health") being the most informative first questions to ask. None of the symptom questions are asked at all, possibly hinting at the redundancy of such questions in language-based assessments for depression severity.

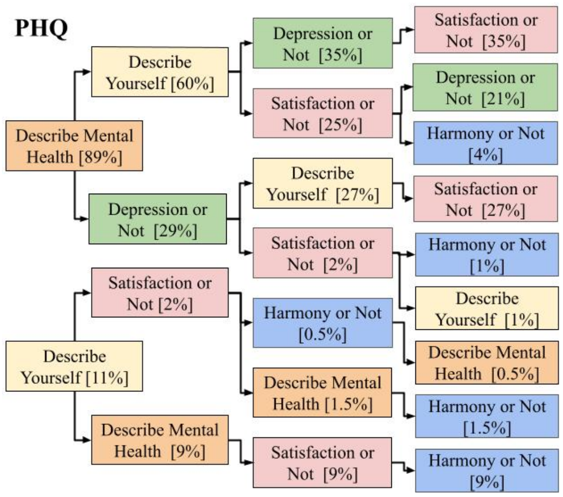

Figure 4: Flowchart of the items picked at n^{th} question using ALIRT. The selections of questions for the first few items is rather sparse. Since the latent variable estimate does achieve a high correlation with the classical psychometric measures in 3-5 questions, it hints at the irrelevance of some questions towards the psychometric measure despite high individual feature correlations.

6 Related Work

Over the past decade, researchers have been exploring techniques for mental health assessment (Coppersmith et al., 2015). Initial studies inspired by leveraging communication in social media, indicated that NLP models could moderately accurately predict self-disclosed mental health conditions or events (Coppersmith et al., 2015; De Choudhury et al., 2016), scores from self-report mental health questionnaires (Schwartz et al., 2014; Chancellor and De Choudhury, 2020), and achieve scores aligned with standard screening surveys when compared to clinical records of depression (Eichstaedt et al., 2018). However, such methods only work well with a fairly active social media usage (Kern et al., 2016). While some have proposed methods to utilize transformers with smaller datasets (Ganesan et al., 2021), such an approach is still limited to those willing to share such data or having any of it at all. Further, it has recently been shown that the accuracy of language-based assessments can reach even greater ($r > 0.8$) when the assessment is based on prompting participants for language responses related to mental health, mirroring standard questionnaires but using language responses instead (Kjell et al., 2022). Still, past work has mostly been validated against summed, or averaged, questionnaire scores, while here we consider improved measurement paradigms that rely on latent variables, such as item-response theory (Reise and Waller, 2009).

Within the domains of NLP, IRT has been used in chatbot evaluation (Sedoc and Ungar, 2020), for textual entailment (Lalor et al., 2016). IRT has also been used to impute missing data (Pliakos et al., 2019) and to compare different ML classifiers at an instance level. Feature/question selection (an NP-complete problem) has also been explored with IRT over a number of fixed selection, ranking, and ordering methods in the recent years (Abdel-Aal and El-Alfy, 2009; Kline et al., 2020; Coban, 2022b). In a related study, (Coban, 2022a) applied IRT to linguistic data, converting language into a term-document matrix for feature selection. However, our approach in adaptive language-based assessment extends beyond fixed feature selection settings. We aim to dynamically adapt to each data sample, facilitated by IRT-based ordering. It's important to note that adaptive language testing differs from personalized recommendation systems. While the latter emphasizes item similarity, language-based testing strives to precisely assess users' latent traits, setting it apart from recommendation systems designed for preferences.

7 Conclusion

Mental health issues vary widely across individuals, suggesting the need for assessments that can enable wide ranging symptoms and be adaptive to the individual. We introduced the task of adaptive language-based assessment for eliciting the most informative responses as well as developed and explored two methods to perform the task, ALIRT and the *Actor-Critic* methods, along with a suite of more straight-forward approaches. Evaluated against depression severity scores derived from 11 questions, ALIRT was able to capture over 90% of the variance explained (R^2) after only 4 questions while optimal fixed ordering approaches needed at least 7, suggesting patient time could be saved with this approach. We further saw that a regression approach that tries to optimally weight question-scores had only minor benefits over the IRT-based ($\hat{L}$), that ALIRT generalized to assessing anxiety in addition to depression, and that symptom-focused questions were not as informative (never chosen early) as compared to broader questions. The adaptive approach, in general, can significantly reduce the number of questions required to achieve high validity, as well as yield insights into the questions that produce the most informative responses suggesting better question/prompt creation.

8 Acknowledgements

We would like to thank the reviewers for their valuable feedback that helped us improving this paper.

This work was supported in part by a grant from the CDC/NIOSH (U01 OH012476), a grant from the NIH-NIAAA (R01 AA028032) and a DARPA Young Faculty Award grant #W911NF-20-1-0306 awarded to H. Andrew Schwartz at Stony Brook University. The conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, NIH, any other government organization, or the U.S. Government.

9 Limitations

This work has a few key limitations: for Classical Test Theory (CTT), we assessed outcomes using self-report questionnaires, specifically PHQ-9 and GAD-7. However, relying on self-reporting in surveys may not ensure complete reliability for diagnostic accuracy. Nevertheless, such self-reported measures have demonstrated consistent associations with diagnoses, proving valuable in clinical assessment and treatment contexts beyond diagnosis (Kroenke et al., 2001). For instance, anxiety scores from self-reported surveys have shown strong correlations with significant real-world outcomes like mortality (Kikkenborg Berg et al., 2014). To validate the assessments proposed in this study, it is crucial to evaluate them against clinical outcomes.

The study was limited by the number of data points and use of descriptive words in English, instead of open-ended texts, due to which we use word embeddings instead of contextual embeddings. While the results in this paper that make the case for adaptive testing should likely translate to other domains including open-ended questions and response domains, we leave that direction open for future work. Instead, we view our work as a first step in integrating adaptive testing into chatbot-style mental health assessments, with a small dataset of descriptive word responses.

10 Ethical Considerations

The dataset used was collected from participants in Prolific and Amazon Mechanical Turk, who were paid to respond to the 11 descriptive questions, along with PHQ-9 and GAD-7 questionnaires. The participants were English-speaking and geographically located in the UK. All of the data is anonymized. The research was approved by an academic institutional ethics review board (exempt status).

This method could potentially be used in the wild– social media posts disclosing diagnoses could be abused to train larger models and track people's latent psychological traits at each utterance in their language, exposing vulnerable people on social media to potential exploitation.

However, as NLP advances in enhancing human-focused applications, such as improving mental health assessment, the balance between considerations for human privacy and open data sharing becomes crucial. In this instance, the data used was shared only with consent for academic research, and open sharing violates trust with participants and ethical review board agreements. Benton et al. (2017) extensively discusses these issues. While the ideal is to release everything while preserving privacy, the limited availability of data suggests an imperative for those with access to share our work openly within ethical guidelines.

References

Radwan E Abdel-Aal and El-Sayed M El-Alfy. 2009. Constructing optimal educational tests using gmdh-based item ranking and selection. *Neurocomputing*, 72(4-6):1184–1197.

Edgar Altszyler, Mariano Sigman, Sidarta Ribeiro, and Diego Fernández Slezak. 2016. Comparative study of lsa vs word2vec embeddings in small corpora: a case study in dreams database. *arXiv preprint arXiv:1610.01520*.

DSMTF American Psychiatric Association, American Psychiatric Association, et al. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*, volume 5. American psychiatric association Washington, DC.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 94–102.

Jason Catlett. 1991. On changing continuous attributes into ordered discrete attributes. In *Machine Learning—EWSL-91: European Working Session on Learning Porto, Portugal, March 6–8, 1991 Proceedings 5*, pages 164–178. Springer.

R Philip Chalmers. 2012. mirt: A multidimensional item response theory package for the r environment. *Journal of statistical Software*, 48:1–29.

R Philip Chalmers. 2016. Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, 71:1–38.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):43.

Onder Coban. 2022a. Irtext: An item response theory-based approach for text categorization. *Arabian Journal for Science and Engineering*, 47(8):9423–9439.

Onder Coban. 2022b. A new modification and application of item response theory-based feature selection for different machine learning tasks. *Concurrency and Computation: Practice and Experience*, 34(26):e7282.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 31–39.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110.

Charles D Dziuban and Edwin C Shirkey. 1974. When is a correlation matrix appropriate for factor analysis? some decision rules. *Psychological bulletin*, 81(6):358.

Johannes C Eichstaedt, Margaret L Kern, David B Yaden, H Andrew Schwartz, Salvatore Giorgi, Gregory Park, Courtney A Hagan, Victoria A Tobolsky, Laura K Smith, Anneke Buffone, et al. 2021. Closed- and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods*, 26(4):398.

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

Adithya V Ganesan, Matthew Matero, Aravind Reddy Ravula, Huy Vu, and H Andrew Schwartz. 2021. Empirical evaluation of pre-trained transformers for human-level nlp: the role of sample size and dimensionality. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2021, page 4515. NIH Public Access.

Richard L Gorsuch. 1973. Using bartlett's significance test to determine the number of factors to extract. *Educational and Psychological Measurement*, 33(2):361–364.

Ivo Grondman, Lucian Busoniu, Gabriel AD Lopes, and Robert Babuska. 2012. A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):1291–1307.

Anders Hald. 1999. On the history of maximum likelihood in relation to inverse probability and least squares. *Statistical Science*, 14(2):214–222.

HF Kaiser, John Rice, Jiffy Little, and I Mark. 1974. Educational and psychological measurement. *American Psychological Association*, 34:111–7.

Margaret L Kern, Gregory Park, Johannes C Eichstaedt, H Andrew Schwartz, Maarten Sap, Laura K Smith, and Lyle H Ungar. 2016. Gaining insights from social media language: Methodologies and challenges. *Psychological methods*, 21(4):507.

Selina Kikkenborg Berg, Lau Caspar Thygesen, Jesper HASTRUP Svendsen, Anne Vinggaard Christensen, and Ann-Dorthe Zwisler. 2014. Anxiety predicts mortality in icd patients: results from the cross-sectional national copenhearticd survey with register follow-up. *Pacing and Clinical Electrophysiology*, 37(12):1641–1650.

Oscar NE Kjell, Katarina Kjell, Danilo Garcia, and Sverker Sikström. 2019. Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods*, 24(1):92.

Oscar NE Kjell, Sverker Sikström, Katarina Kjell, and H Andrew Schwartz. 2022. Natural language analyzed with ai-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific reports*, 12(1):3918.

Adrienne Kline, Theresa Kline, Zahra Shakeri Hossein Abad, and Joon Lee. 2020. Novel feature selection for artificial intelligence using item response theory for mortality prediction. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5729–5732. IEEE.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

John P Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 648. NIH Public Access.

Dong C Liu and Jorge Nocedal. 1989. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528.

Frederic M Lord. 2012. *Applications of item response theory to practical testing problems*. Routledge.

Frederic M Lord and Melvin R Novick. 2008. *Statistical theories of mental test scores*. IAP.

David N Milne, Glen Pink, Ben Hachey, and Rafael A Calvo. 2016. Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the third workshop on computational linguistics and clinical psychology*, pages 118–127.

Eiji Muraki and Mari Muraki. 2016. Generalized partial credit model. In *Handbook of item response theory*, pages 155–166. Chapman and Hall/CRC.

Remo Ostini and Michael L Nering. 2006. *Polytomous item response theory models*. 144. Sage.

Konstantinos Pliakos, Seang-Hwane Joo, Jung Yeon Park, Frederik Cornillie, Celine Vens, and Wim Van den Noortgate. 2019. Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems. *Computers & Education*, 137:91–103.

Steven P Reise and Niels G Waller. 2009. Item response theory and clinical measurement. *Annual review of clinical psychology*, 5:27–48.

Fumiko Samejima. 2016. Graded response models. In *Handbook of item response theory*, pages 123–136. Chapman and Hall/CRC.

H Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 118–125.

João Sedoc and Lyle Ungar. 2020. Item response theory for efficient human evaluation of chatbots. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 21–33, Online. Association for Computational Linguistics.

Sverker Sikström, Alfred Pålsson Höök, and Oscar Kjell. 2023. Precise language responses versus easy rating scales—comparing respondents' views with clinicians' belief of the respondent's views. *Plos one*, 18(2):e0267995.

Marcin Siwek, Dominika Dudek, Janusz Rybakowski, Dorota Łojko, Tomasz Pawłowski, and Andrzej Kiejna. 2009. Mood disorder questionnaire–characteristic and indications. *Psychiatria Polska*, 43(3):287–299.

Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine*, 166(10):1092–1097.

Yoshio Takane and Jan De Leeuw. 1987. On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52(3):393–408.

Lingling Xu, Ruyi Jin, Feifei Huang, Yanhui Zhou, Zonglong Li, and Minqiang Zhang. 2020. Development of computerized adaptive testing for emotion regulation. *Frontiers in Psychology*, 11:561358.

A Dimension Reduction on Contextual Embeddings

Our experiment was limited to static embeddings trained specifically on the mental health domain, which was due to the data scarcity and format (descriptive words), leading to the need for a small embedding size. For comparability and to show the utility of adaptive testing with other embeddings, we experiment with dimension reduction on RoBERTA-large model. We separately collect sets of five words describing daily emotions, mood, feelings etc. from 572 users, for about 30 days each. RoBERTA-large (1024 dim) embeddings are extracted for each of these sets, on which PCA is applied to reduce the dimensions from 1024 to 10. The reduction was then applied to all the questions in Table 3, and then trained ALIRT and *Actor-Critic* models. The results are reported in Table A1. The same procedure was repeated with GloVe embeddings (cased, trained on Common Crawl) by learning a reduction on 300 dimensional GloVe vectors to 10 dimensions for fair comparison. The results for GloVe are reported in A2.

Method	Eval. against	1	2	4
FixedFor-$\widehat{L}$	CTT	0.582	0.653	0.708
FixedFor-$\widehat{Y}$	CTT	0.591	0.670	**0.718**
ActorCritic-$\hat{L}$	CTT	0.575	0.652	0.682
ActorCritic-$\widehat{Y}$	CTT	**0.604**	**0.678**	0.705
ALIRT-$\hat{L}$	CTT	0.592	0.651	0.706
ALIRT-$\widehat{Y}$	CTT	0.596	0.661	0.715
FixedFor-$\hat{L}$	$\hat{L}_{all}$	0.762	0.857	0.934
FixedFor-$\widehat{Y}$	$\hat{L}_{all}$	0.773	0.864	0.929
ActorCritic-$\hat{L}$	$\hat{L}_{all}$	0.740	0.828	0.913
ActorCritic-$\widehat{Y}$	$\hat{L}_{all}$	0.769	0.840	0.908
ALIRT-$\hat{L}$	$\hat{L}_{all}$	0.784	0.873	**0.944**
ALIRT-$\widehat{Y}$	$\hat{L}_{all}$	**0.792**	**0.878**	0.938

Table A1: Comparison of fixed and adaptive strategies with 10-dimensional contexutual embeddings reduced from RoBERTA-large, evaluated against PHQ-9 for CTT and $\hat{L}_{all}$– the latent score derived when using all the items – for IRT .

ALIRT is a better choice when using RoBERTA-large as well, especially when using IRT scoring strategy, but does not compromise much on the performance given the number of parameters in Classical Test Theory too. Forward selection is comparable to adaptive testing among fixed ordering methods. However, the difference between fixed and adaptive strategies is not as significant as when using static embeddings. This can be explained with the context-independent word responses in the dataset used, where contextual embeddings do not seem to improve the predictive power.

Method	Eval. against	1	2	4
FixedFor-$\widehat{L}$	CTT	0.626	0.703	0.732
FixedFor-$\widehat{Y}$	CTT	0.637	0.714	0.747
ActorCritic-$\hat{L}$	CTT	0.605	0.695	0.729
ActorCritic-$\widehat{Y}$	CTT	0.628	**0.723**	**0.750**
ALIRT-$\hat{L}$	CTT	0.630	0.660	0.719
ALIRT-$\widehat{Y}$	CTT	**0.644**	0.712	0.748

Table A2: Comparison of fixed and adaptive strategies with 10-dimensional word embeddings that were reduced with GloVe embeddings, evaluated against PHQ-9 for CTT. Consistent with the results observed with LSA and RoBERTA-large embeddings, the adaptive methods perform better than fixed. Further, the effect observed with GloVe is comparable to that of LSA as opposed to RoBERTA-large since it is non-contextual and better suited for descriptive words rather than open-ended language.

B Computational Complexity of the *Actor-critic* model

For N items, there are $2^N - 1$ combinations of items, and therefore, $2^N - 1$ error prediction models. For N total questions and k questions administered so far, the number of combinations of questions left is $\binom{N}{N-k} = \binom{N}{k}$. Number of items that could be picked next is $(N-k)$ Adding them over all the possibilities:

$$(N-1).\binom{N}{1} + ... + (N-N).\binom{N}{N}$$

$$= \sum_{1}^{N}(N-k).\binom{N}{k} = -N + \sum_{0}^{N}(N-k).\binom{N}{k}$$

$$= -N + N\sum_{0}^{N}\binom{N-1}{k} = N(2^{N-1} - 1)$$

We arrive at a complexity of $O(N.2^N)$.

C Dataset Dimensionality

Item response theory is a form of factor analysis (Takane and De Leeuw, 1987). Therefore, we perform two tests to ensure the feasibility of our dataset. (Dziuban and Shirkey, 1974) Kaiser–Meyer–Olkin (KMO) test (Kaiser et al., 1974) checks sampling adequacy for each feature based on the correlation matrix and produces a KMO value between 0-1. The higher the KMO value is, the better suited the data is for factor analysis. Our dataset has a KMO value of 0.924, which makes it highly suitable for factor analysis. We also perform the Bartlett Test of Sphericity on our dataset to determine the number of significant factors. (Gorsuch, 1973) The test results in a p-value $< .001$, which indicates that the IRT latent variable should indeed capture the features, with the Kaiser criterion indicating there is just 1 latent factor.

D IRT parameters for ALIRT

The polytomous model fits a 2-parameter (2PL) characteristic curve for each polytomous threshold for each item. 2PL item characteristic curve is typically modeled with two parameters:

$$P(\theta) = \frac{1}{1 + e^{-\alpha(\theta - \beta)}}$$

where β is the difficulty parameter (midpoint of the slope; models how “difficult” an item is) and α is the discriminant (slope of the midpoint; it models how well an item discriminates between participants that score higher/lower than the difficulty). For polytomous IRT modeling, if the responses are polytomized to K values [0,1, ... K-1, K], then there are K-1 logistic characteristic curves learned for each threshold: between 0 and 1, between 1 and 2 ... and between K-1 and K. In our case, a single discriminant α is learned across all the K-1 curves per item. Therefore, for jth item and kth curve, the item characteristic function is:

$$P(\theta_k^j) = \frac{1}{1 + e^{-\alpha^j(\theta - \beta_k^j)}} - \frac{1}{1 + e^{-\alpha^j(\theta - \beta_{k-1}^j)}}$$

The total number of parameters for J questions, with K-tomous responses is therefore J x K. Maximum Fisher Information (MFI) is the objective used by ALIRT to pick the next best question. This is calculated as the derivative of log probabilities at the current latent estimate using the item characteristic functions. (Hald, 1999) MFI picks the question with highest variance in the estimate of the score/latent variable. The latent variable is clipped between -6 and +6.