USING IRT TO ANALYZE A PSYCHOMETRIC SURVEY ON DEPRESSION

Daniel Valverde & Zoe Berling

3/13/2022

USING IRT TO ANALYZE A PSYCHOMETRIC SURVEY ON DEPRESSION

Executive summary/ Abstract:

Introduction:

In our study we sought out to see how well the survey measures anxiety and depression. We fit an Item Response Theory model to data from a psychometric survey measuring depression to optimize the survey.

Data, Data Analysis, and Cleaning:

Our dataset came from a survey that was performed on approximately 40,000 participants with the intention of measuring their levels of anxiety and depression. Anyone had access to the survey, and people self-selected to take it to get personalized results. The dataset included 42 questions and tracked the participant response, the amount of time it took them to respond, and the order the question appeared in the survey. We chose to focus on the first six questions of the survey and only focused on the participants' responses.

Implementation:

For our implementation, we utilized the LTM library to implement a dual parameter Item Response Theory Analysis on our dataset. Our analysis focused on the difficulty level and distinguishing capacity of a subset of questions from the survey. We went through each of the assumptions of IRT which are: monotonicity, one-dimensionality, item independence, and item invariance. We utilized Item Characteristic curves to better visualize the difficulty and discrimination of each question for participants with different levels of depression.

Findings and Conclusion:

We discovered that for the most part questions 1 through 6 on the survey were all adequate but question 2 performed worse than the others at measuring depression in the participants. To improve our survey, we could improve by either removing question 2 ("I was aware of dryness of my mouth") or potentially changing it to be something more meaningful that would measure participants' levels of depression better. We also discovered that men and women tend to respond differently to question 3 ("I couldn't seem to experience any positive feelings at all") and question 5 ("I just couldn't seem to get going.")

Dataset:

We chose a Kaggle dataset[1] of a depression and anxiety survey with ~40k responses. The psychometric portion of the survey includes 42 questions with the following scale from 1-4:

1 = Did not apply to me at all

2 = Applied to me to some degree, or some of the time

3 = Applied to me to a considerable degree, or a good part of the time

4 = Applied to me very much, or most of the time

In addition to measuring the participant's answer to the question, the amount of time the participant took to mark their response and the order the question appeared in the survey was also saved in the dataset. The survey also included a generic demographics questionnaire at the end, as well as a Ten Item Personality Inventory[2]. Finally, the survey ended with a checklist with the instructions: "In the grid below, check all the words whose definitions you are sure you know".

We wanted to know the efficacy of the survey to measure a participant's latent depression and were specifically interested to know if there were any questions that were not good indicators and could potentially be cut out of the survey.

Item Response Theory (IRT) is a family of statistical models fit to survey data that attempts to explain relationships between latent traits and their observable manifestations. The classical case for IRT is in the educational domain to test the efficacy of a survey in measuring someone's intelligence, but there are many use cases for IRT to optimize psychometric surveys and it is a great fit to analyze our chosen dataset.

Some use-cases for IRT include but are not limited to ensuring that if the results of a survey are the same across different demographic groups, comparing two different tests that measure the same latent trait, and understanding which questions are the best and which can be cut out of surveys to optimize survey length.

[1] https://www.kaggle.com/yamqwe/depression-anxiety-stress-scales
[2] http://gosling.psy.utexas.edu/scales-weve-developed/ten-item-personality-measure-tipi/

The type of IRT model depends on the scale of the survey. Dichotomous IRT methods are applied to surveys with binary responses, while polytomous IRT methods can be applied to surveys with more than two response variables.

For the scope of this project, we chose a subset of six questions and dichotomized the responses, changing the scale from 1-4 to binary to apply a dichotomous IRT model. We only selected data columns for the response and the gender of the participant.

Assumptions and Implementation of IRT:

There are a few libraries that could prove handy for implementing IRT. Two common ones are LTM and MIRT.

Depending on the logistic model you want to use for IRT you would choose between a single parameter, a dual parameter, or a triple parameter model. The single parameter only calculates the item difficulty level (b_i). The dual parameter model calculates the item difficulty (b_i) and the item distinguishing capacity (a_i), and the guessing parameter (c_i) is set to zero. Lastly, the triple parameter model calculates all 3 parameters: difficulty, distinguishing capacity, and guessing parameter.

The model that bests fits our data is a two-parameter logistic (2PL) IRT model that returns item difficulty and item discrimination.

The following formula of a two-parameter logistic (2PL) IRT models the student's ability (θ):

$$P(X = 1|\theta, a, b) = \frac{e^{a(\theta - b)}}{1 + e^{a(\theta - b)}}$$

This logistic equation[3] returns an Item Characteristic Curve (ICC) that describes the probability that an individual with latent ability level θ scores highly on a given item (X=1) with two item characteristic parameters.

The item difficulty, b, is the parameter that is used to describe how likely an item is to achieve a .5 probability of a high response at a given level of the latent trait. The item discriminability, b, is how well an item is able to discriminate against different levels of θ (the latent trait).

In an ICC graph, item difficulty is mapped along the x-axis (latent trait ability). The item calibration is where the slope crosses the y axis at .5 probability and measures the

[3] https://www.publichealth.columbia.edu/research/population-health-methods/item-response-theory

amount of the latent trait needed to score high on a question. If the curve reaches .5 probability at a lower x value, a person would need to exhibit less of the latent trait to achieve a high score, while if the curve reaches .5 at a higher x value, a person would likely exhibit more of the latent trait. Item discriminability is reflected in the steepness of the slope of the item characteristic curves. The steeper the slope, the higher rate at which the probability of scoring high on an item increases given the participant's latent ability level. A more gradual slope indicates the item is not as good at discriminating between people with various levels of the latent trait. **(see Figure 0)**.

Within our analysis, difficulty measures the likelihood an item has to return a 50% probability at a given level of depression[4]. Within the output, item difficulty is essentially a z-score for the latent variable. Looking at our dual parameter IRT model showed that of the questions we analyzed every question was within one standard deviation of the mean when focusing on the difficulty output **(see Figure 1)**.

Discrimination on the other hand can be calculated by the slope and determines how well each question measures the participants. When analyzing the results of the discrimination parameter, numbers between 1 and 4 are typically considered to be within a good range. The higher the better. In a comparison between the ICC curves for our item set, you can see that the results of running our IRT model showed that questions 1, 3, and 5 were better at distinguishing above-average levels of depression; while question 2 although not being terrible at distinguishing levels of depression was not quite as good as the others **(see Figure 1)**.

Chi-squared[5] is a common method used to measure an IRT model's item fit. When it comes to Item Response Theory you typically want your data points to be independent but when measuring this with chi-squared, due to the nature of having a large dataset, everything ends up being dependent. Therefore, it is typically common practice not give to much weight to the results of chi-squared. **(see Figure 2)**.

The Item Characteristic curves for our model satisfy the assumption of monotonicity, given that the shape of our curves for all items in our LTM model have a general S shape. When analyzing difficulty within our model we compare where each curve crosses the line at .5 probability. In our analysis, we can see that a lower level of depression is required to score highly on Question 1 while for Question 4 a higher level of depression is required to score highly **(see Figure 3)**.

[4] https://files.eric.ed.gov/fulltext/EJ1133065.pdf (pp.259)
[5] https://www.youtube.com/watch?v=VtWsyUCGfhg

The item information curve displays the first derivative of the item characteristic curve. Dependent on where the center of the bell curve lies relates to which part of the population that curve is better at discerning.

Our item information curve emphasizes which questions are more informative for different levels of depression. Q1 is very good at distinguishing scorers with average depression. While Q3 is good at distinguishing scorers with slightly higher levels of depression, due to the slight shift to the right **(see Figure 4)**.

To check the assumption of invariability, we separated the data into responses from men and women and graphed the respective ICC curves side by side **(see Figure 5)**. For the most part, the curves are parallel, however, the lines for men and women in items 3 and 5 intersect, indicating there is a difference in the way men and women answer this question. Since the slopes and item calibration are very similar, we can conclude that the item parameters still behave similarly in these different populations[6].

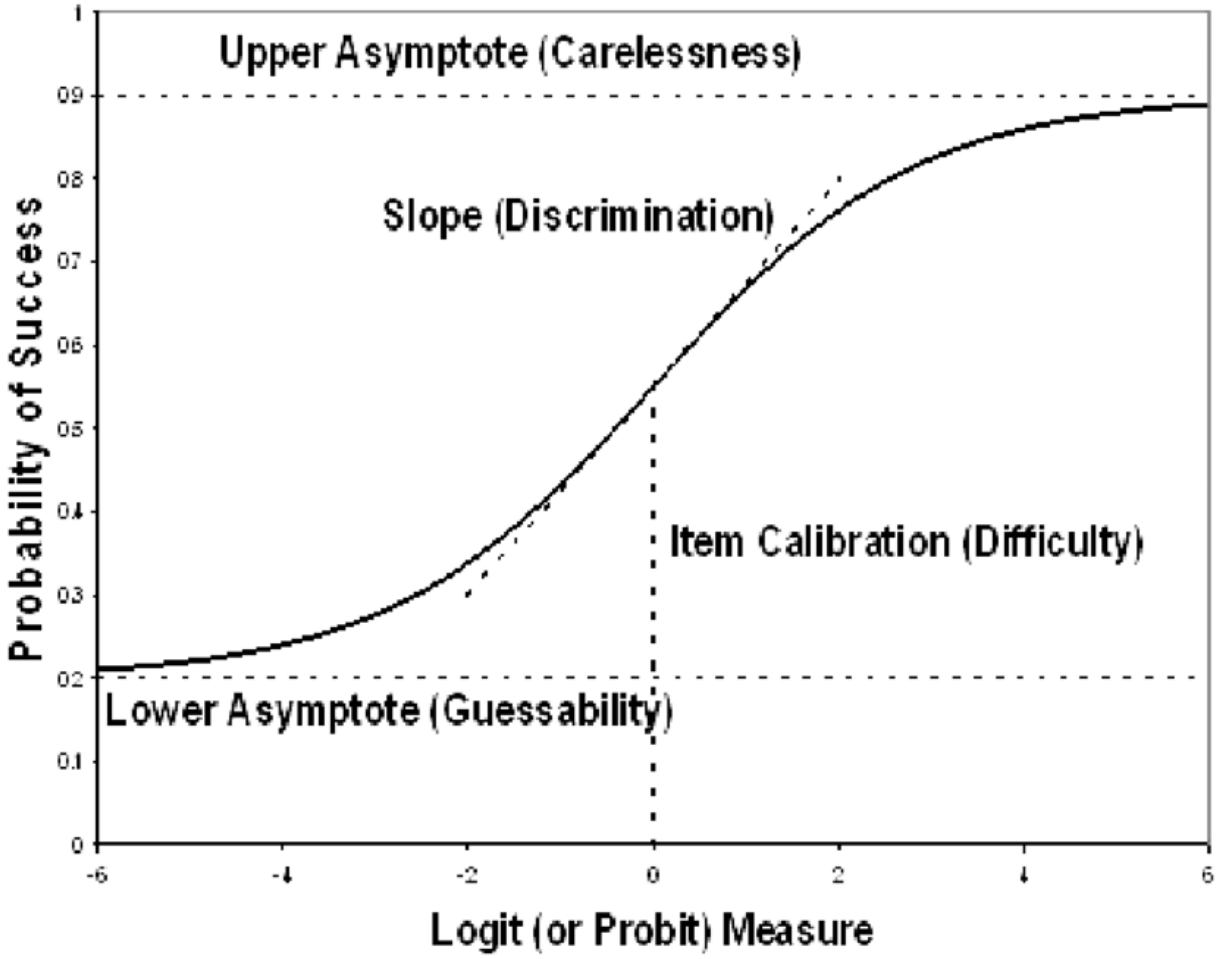
The factor scores method within the LTM library outputs every combination of responses from the survey. Factor scores give the pattern of data and what one would expect the z-score to be given that pattern of data. In the “Obs” column it can be noted that responding no to everything contained the most observations while responding yes to everything had the second most observations. The “Exp” column displays the expected value of the observations column based on the dataset. The “z1” is the expected ability score for each combination of responses. **(see Figures 6A, 6B, 6C)**

Conclusion:

The question we sought out to answer was: If someone is depressed, how good is our survey at measuring this? Through our analysis, we determined that most of the questions in our selected subset were adequate at measuring latent depression. In order to improve our survey, we would either remove question 2 or try changing it to be something more meaningful that would potentially measure participants’ levels of depression better. While our analysis was limited by the scope of our project, we got a glimpse into the potential applications of IRT. Future investigations could involve fitting more questions into the model, running a polytomous IRT analysis on the full survey scale, or even factoring in other variables saved in the original data set, such as the amount of time it took participants to respond, or responses to the Ten Item Personality Inventory. IRT is a great investigative method to explore psychometric surveys such as the one we chose.

[6] https://aidenloe.github.io/introToIRT.html

(Figure 0)



[7]

[7] https://www.researchgate.net/figure/A-typical-item-characteristic-curve-ICC_fig2_315785883

(Figure 1)

```
model<-ltm(dat_base~z1, IRT.param=TRUE)
print(coef(model))
```

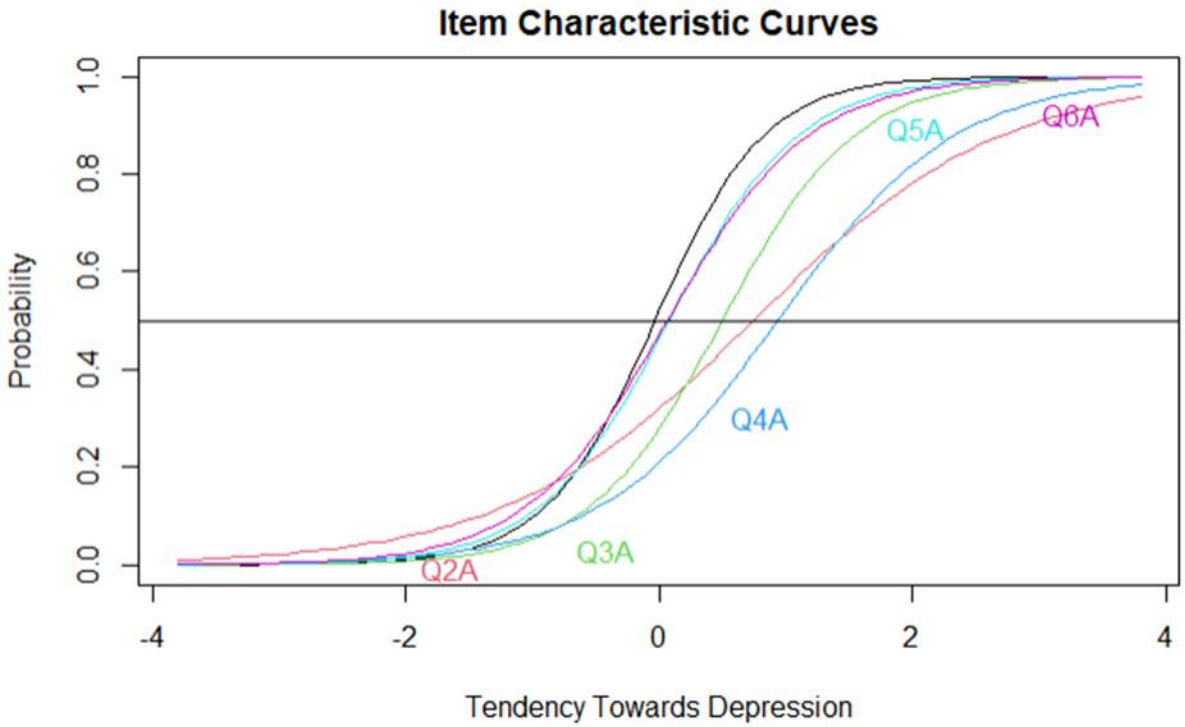
```
        Dffclt   Dscrmn
Q1A -0.03346154 2.317070
Q2A  0.74285394 1.016176
Q3A  0.49802722 1.908842
Q4A  0.93501859 1.423853
Q5A  0.07939943 1.953463
Q6A  0.06521985 1.790820
```

(Figure 2)

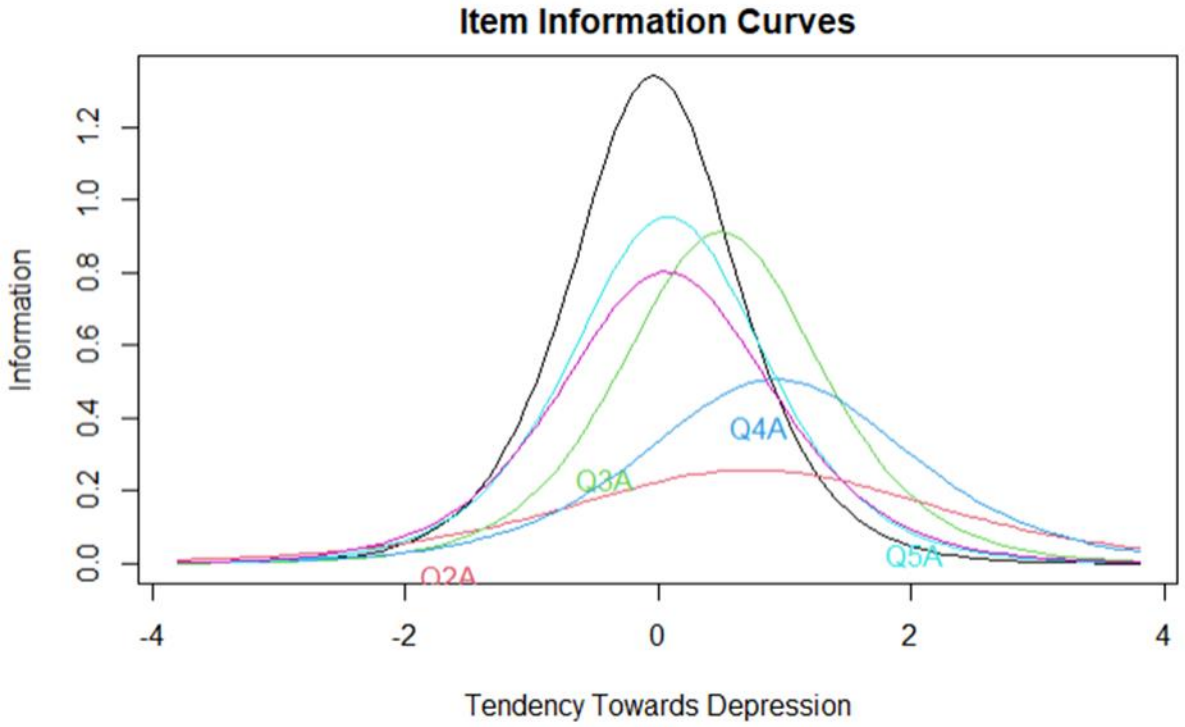
	X^2 <dbl>	Pr(>X^2) <chr>
Q1A	3789.449	<0.0001
Q2A	9243.511	<0.0001
Q3A	1704.668	<0.0001
Q4A	3804.557	<0.0001
Q5A	2943.172	<0.0001
Q6A	1302.295	<0.0001

6 rows

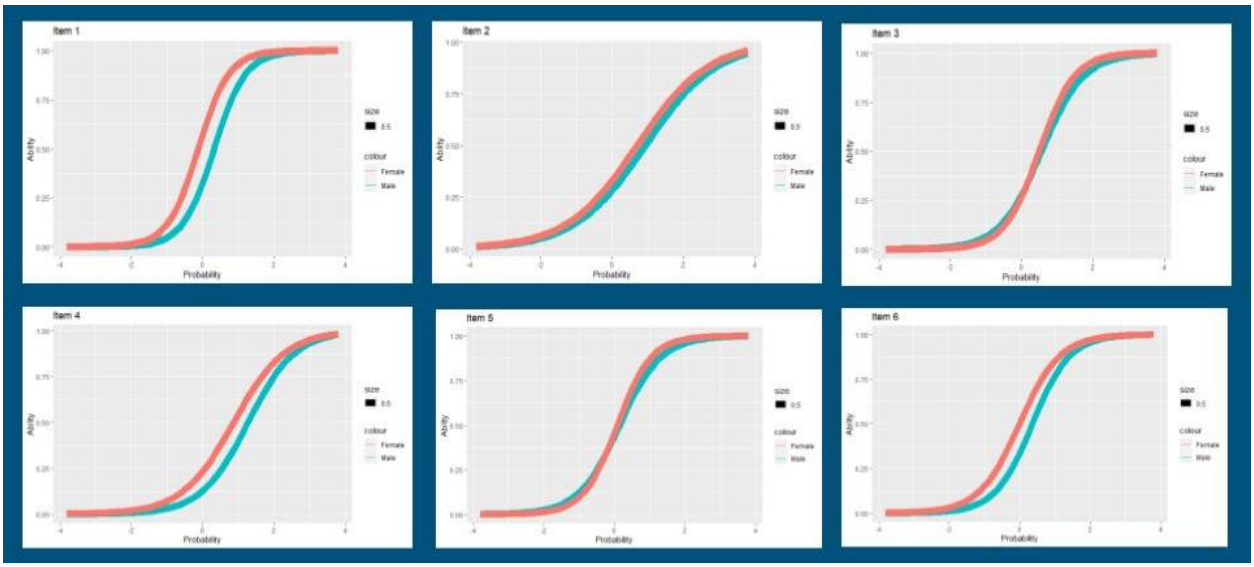
(Figure 3)


Item Characteristic Curves
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Tendency Towards Depression
Q2A
Q3A
Q4A
Q5A
Q6A

(Figure 4)



(Figure 5)



(Figure 6A)

Q1A <dbl>	Q2A <dbl>	Q3A <dbl>	Q4A <dbl>	Q5A <dbl>	Q6A <dbl>	Obs <dbl>	Exp <dbl>	z1 <dbl>	se.z1 <dbl>
0	0	0	0	0	0	8336	8061.059	-1.001	0.617
0	0	0	0	0	1	1421	1595.205	-0.472	0.486
0	0	0	0	1	0	1407	1406.395	-0.435	0.479
0	0	0	0	1	1	514	733.086	-0.067	0.435
0	0	0	1	0	0	478	592.888	-0.562	0.504

(Figure 6B)

Q1A <dbl>	Q2A <dbl>	Q3A <dbl>	Q4A <dbl>	Q5A <dbl>	Q6A <dbl>	Obs <dbl>	Exp <dbl>	z1 <dbl>	se.z1 <dbl>
0	1	0	1	0	0	262	163.498	-0.327	0.462
0	1	0	1	0	1	190	103.961	0.024	0.430
0	1	0	1	1	0	128	100.407	0.054	0.429
0	1	0	1	1	1	160	133.124	0.385	0.436
0	1	1	0	0	0	142	203.466	-0.226	0.449
0	1	1	0	0	1	98	156.215	0.113	0.428
0	1	1	0	1	0	312	153.332	0.143	0.428
0	1	1	0	1	1	180	243.723	0.478	0.442
0	1	1	1	0	0	43	45.249	0.046	0.429
0	1	1	1	0	1	57	59.007	0.376	0.435

(Figure 6C)

Q1A <dbl>	Q2A <dbl>	Q3A <dbl>	Q4A <dbl>	Q5A <dbl>	Q6A <dbl>	Obs <dbl>	Exp <dbl>	z1 <dbl>	se.z1 <dbl>
0	1	0	1	0	0	262	163.498	-0.327	0.462
0	1	0	1	0	1	190	103.961	0.024	0.430
0	1	0	1	1	0	128	100.407	0.054	0.429
0	1	0	1	1	1	160	133.124	0.385	0.436
0	1	1	0	0	0	142	203.466	-0.226	0.449
0	1	1	0	0	1	98	156.215	0.113	0.428
0	1	1	0	1	0	312	153.332	0.143	0.428
0	1	1	0	1	1	180	243.723	0.478	0.442
0	1	1	1	0	0	43	45.249	0.046	0.429
0	1	1	1	0	1	57	59.007	0.376	0.435