An Item Response Theory Analysis of the Patient Health Questionnaire

Jamie C. Chiu[1, 2]

[1] Department of Psychology, Princeton University

[2] Princeton Neuroscience Institute, Princeton University

Author Note

Correspondence concerning this article should be addressed to Jamie C. Chiu, Princeton Neuroscience Institute, Princeton University, Princeton, New Jersey 08540. E-mail: jamiechiu@princeton.edu

Abstract

The 9-item Patient Health Questionnaire (PHQ-9) is a commonly-used depression measurement tool. The ratings on each question is often treated as continuous, summed to a total score, and then compared against different cut-offs to determine a respondent's depression severity. However, the PHQ-9 is an ordinal scale where each question has four rank-ordered responses - "Not at all", "Several days", "More than half the days", and "Almost every day". The following report is an exploration of using an ordinal-appropriate model to analyse the PHQ-9. More specifically, the Graded Response Model from Item Response Theory is applied to examine the psychometric properties of the PHQ-9. The data comprises 2,177 responses to the PHQ-9, gathered from two open data sets. Overall, most of the items of the PHQ-9 has adequate discrimination and difficulty. Item 9 performed the worst, even participants with high depression did not always endorse it. Item 9 may be indicative of a related but separate latent trait instead of depression, and could be potentially excluded from the Patient Health Questionnaire. All analyses were carried out in R and the data and code are provided for reproducibility.

Keywords: item response theory, graded response model, ordinal model, PHQ-9, depression

An Item Response Theory Analysis of the Patient Health Questionnaire

Introduction

The Patient Health Questionnaire (PHQ-9) is a commonly-used 9-item rating scale used to measure depression severity (Kroenke & Spitzer, 2002). The PHQ-9 is an ordinal scale that is often used as a continuous or interval scale – that is, the responses are treated as if the distance between each option is equal. In the PHQ-9, each question asks how much you have experienced a particular symptom over the last 14 days, with rank-ordered responses being "Not at all", "Several days", "More than half the days", and "Almost every day". The ordinal nature of the response categories mean that distances between each response cannot be assumed to be equal. For example, the difference between "Not at all" and "Several days" may be much smaller in the respondent's mind than the difference between "Several days" and "More than half the days". Treating ordinal data as metric by summing across responses and computing a total score can be problematic (Liddell & Kruschke, 2018). (The reader is advised to see Liddell and Kruschke (2018) for an in-depth analysis into the errors and problems that can arise from treating ordinal data as metric.)

One alternative approach for analyzing ordinal rating scales such as the PHQ-9 that involves more than 2 categories of responses is the graded response model (Samejima, 1997) from item response theory. Item response theory is a family of models that aims to look at the underlying latent traits which are driving test performance; and the graded response model is one such model that deals with ordered polytomous categories. This report outlines the analysis of the PHQ-9 scale using a graded response model.

The structure of this report is as follows: Using a real-world data set of PHQ-9 responses, assumptions for item response theory are explored and reported; then a graded response model is fit to the data set and the results are discussed. The data set and code are provided for reproducibility.

Methods

Participants

2,177 participants' responses to the PHQ-9 scale was obtained and combined from two data sets: the Brighten Study (Pratap et al., 2022) and the PERLA Project (Arrabales, 2020). For the purpose of this report, only the PHQ-9 responses were used for analysis, and no participant demographic data was used. For ease of analysis, the PHQ-9 responses are stored as both ordinal (factored and ordered) as well as numerical. To access the data set:

```
PHQ <- read_csv("PHQ9_data.csv")
# head(phq)
```

Table 1

Responses are duplicated in ordinal and numerical values for ease of analysis.

ID	Item 1 (num)	Item 2 (num)	Item 3 (num)	Item 1	Item 2	Item 3
1	0	1	0	Not at all	Several days	Not at all
2	0	1	3	Not at all	Several days	Almost every day
3	1	1	2	Several days	Several days	More than half
4	0	0	1	Not at all	Not at all	Several days
5	0	0	0	Not at all	Not at all	Not at all

Material

Patient Health Questionnaire. The PHQ-9 is a 9-item rating scale that assesses depression symptoms (Kroenke & Spitzer, 2002). To view the items, please refer to the Appendix.

Results

All data analyses were conducted using R (Version 4.2.1; R Core Team, 2022) and the R-packages *FactoMineR* (Version 2.6; Lê, Josse, & Husson, 2008), *lavaan* (Version 0.6.12; Rosseel, 2012), *ltm* (Version 1.2.0; Rizopoulos, 2006), *mokken* (Version 3.0.6; Van der Ark, 2007, 2012), and *psych* (Version 2.2.9; Revelle, 2022). Code chunks are displayed in-line where applicable; for full code to reproduce analyses, please refer to the Appendix. Code for plots are not displayed in-line due to space, please refer to the Appendix for code to reproduce plots.

Examining Fit using Metric Models

Using the numerical values of the PHQ-9 ratings, a normal distribution was fit over each item and the total sum (Figure 1 and 2).

Examining Fit using Item Response Theory

Assumptions Testing. Item response theory models must meet the following assumptions: unidimensionality, local independence, and monotonicity. Each of the assumptions will be discussed below and the statistical methods used to test for them described.

Unidimensionality. Unidimensionality assumes the items all assess the same one underlying latent trait variable. Unidimensionality was tested using principal component analysis to check if only one principal could be sufficiently extracted.

A principal component analysis revealed that 1 component explained about 50% of the variance. Examining the eigenvalues, the first component was the only eigenvalue >1, which is commonly used as a cutoff point for which components are retained. Thus, it can be argued that the assumption of unidimensionality is sufficiently met.

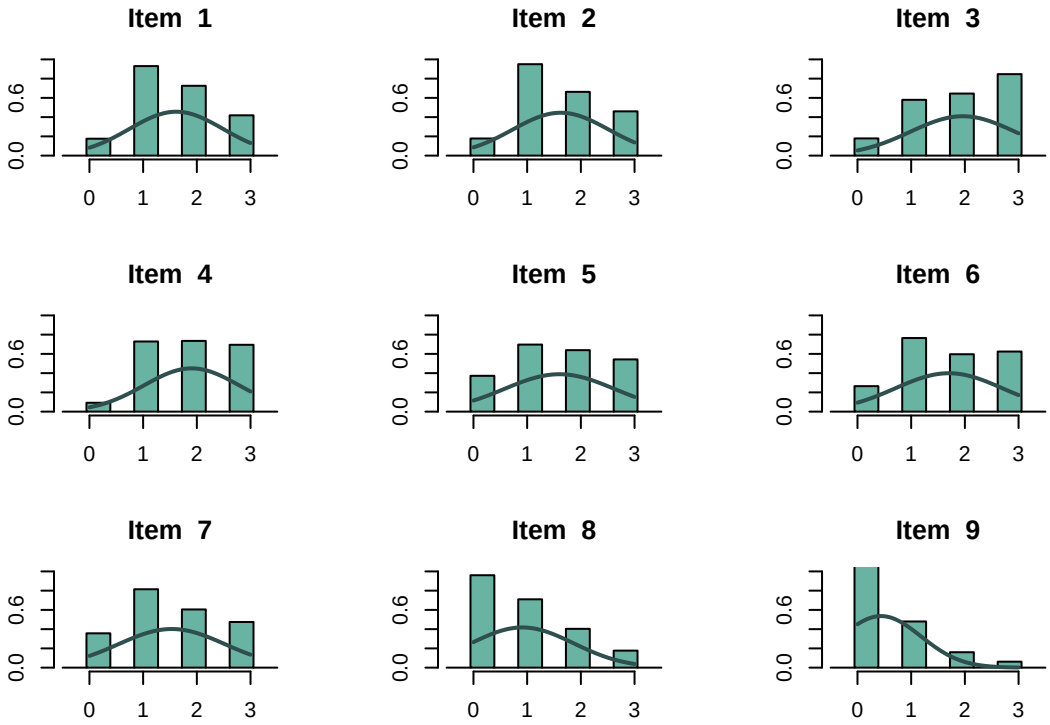

Figure 1. Ratings data from the PHQ-9 items are shown as histograms. Normal distributions from the metric model are superimposed on the data.

However, examining the contribution of variables (Figure 4), it appears that Item 9, which asks about thoughts of wanting to die, loads higher onto a different dimension than the rest of the items.

Local Independence. Local independence assumes the items are only related to the latent trait variable being measured and not to any other factors. Local independence was evaluated using a single factor confirmatory factor analysis and examining the residual correlation matrix. The root mean square error of approximation (RMSEA) was revealed to be 0.1. The cutoff for calculating the probability of a close fit is commonly suggested at 0.1. The comparative fit index (CFI) was shown to be 0.9. CFI assesses the relative improvement in fit of the model compared with the baseline model, with a suggest cutoff of 0.9. Thus, the assumption of local independence is sufficiently met. (For details on the analysis, please refer to the Appendix.)

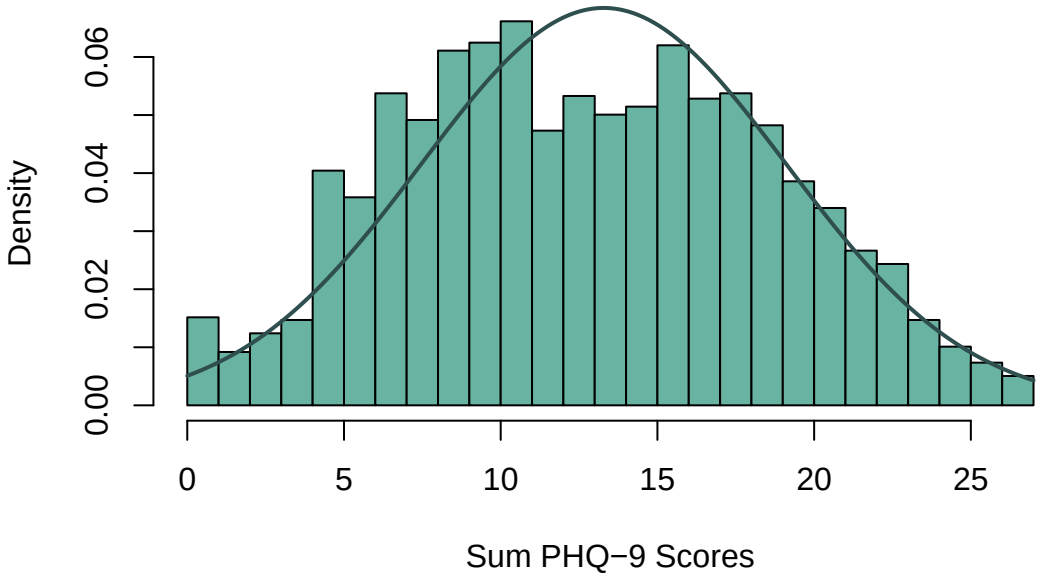

Figure 2. Sum ratings from the PHQ-9 items are shown as a histogram. A metric model normal distribution is superimposed over the data. The fit is poor, as the data histogram bars protrude above and below the normal distribution.

Monotonicity. Monotonicity assumes that the probability of endorsing higher-ranked responses to the items correlates with increasing levels of the latent trait variable (e.g. the probability of choosing "Almost every day" instead of "Several days" should correspond to an increasing level of depression). Monotonicity was tested using a Mokken scaling technique. The fit of the mokken model was evaluated by calculating the scalability coefficient H per item - monotonicity was considered acceptable as the scalability coefficients for each item was >0.30, which is the suggested cutoff.

Graded Response Model. Two graded response models - one with constrained parameters where one discrimination parameter was fixed across all times, and one with unconstrained parameters - was fitted to the data.

Table 2

Principal Component Table of Eigenvalues.

	Eigenvalue	% Variance Explained	Cumulative Variance
comp 1	4.39	48.74	48.74
comp 2	0.92	10.20	58.94
comp 3	0.82	9.08	68.02
comp 4	0.64	7.12	75.14
comp 5	0.54	6.00	81.14
comp 6	0.49	5.39	86.53
comp 7	0.48	5.30	91.83
comp 8	0.43	4.80	96.63
comp 9	0.30	3.37	100.00

```
# fit GRM model
library(ltm)
PHQ_ordinal <- PHQ[, 12:20] # subset ordinal data
# constrained, i.e. discrimination parameter is held constant
mod1 <- grm(PHQ_ordinal, constrained = TRUE)
mod1
# non constrained model
mod2 <- grm(PHQ_ordinal)
mod2
# compare fit
anova(mod1, mod2)
```

The unconstrained model fared better even when accounting for additional parameters ($p < 0.01$). Thus, we will report on the unconstrained model from hereon.

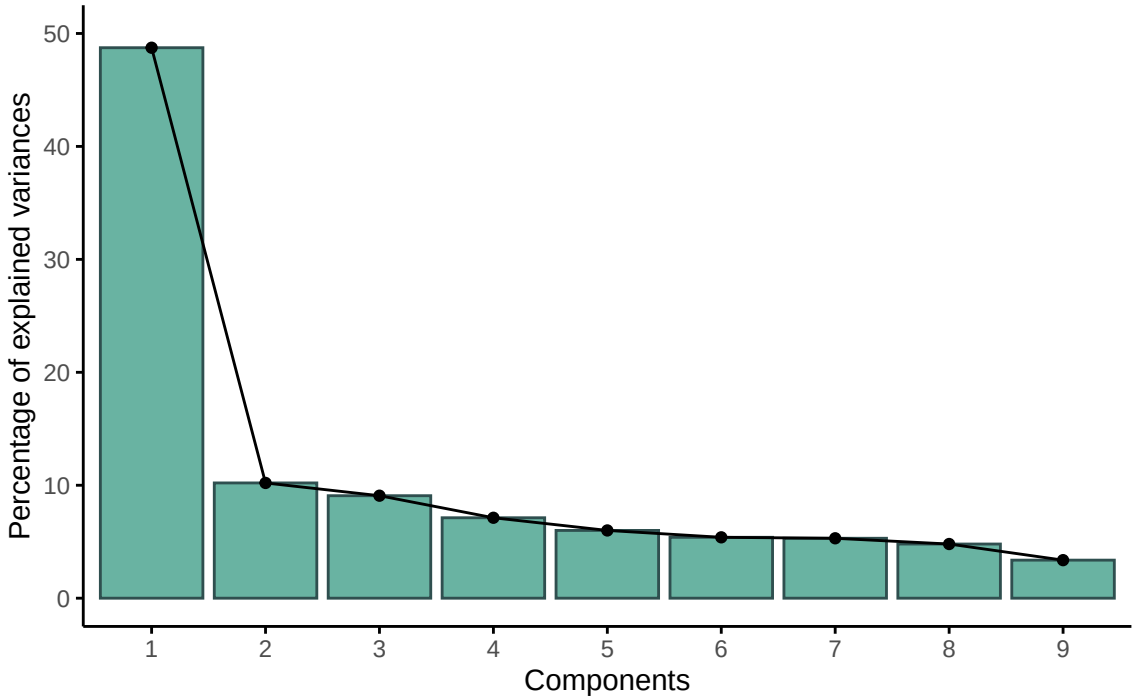

Figure 3. Scree plot of principal component analysis. The first component explains almost 50% of the variance.

Table 4 displays the item thresholds and item discrimination parameters.

Item Threshold. The PHQ-9 has 4 categories for responses, and so there are three threshold parameters: The first threshold indicates the latent trait variable measurement at which there is 50% probability that a respondent would endorse category 1 vs 2. For example, with Item 1, lack of interest, there is a 50% probability that a respondent would endorse either "Not at all" or "Several days" at -1.2 on the latent variable scale. On the other hand, with Item 6, feeling bad about oneself, the 50% probability of endorsing "Not at all" or "Several days" is at -0.7, which indicates a higher latent trait. The second threshold indicates the 50% probability of endorsing category 1 or 2 vs category 3; and the last threshold is category 1,2,3 vs category 4.

The item threshold parameter is also often referred to as the item difficulty – where a

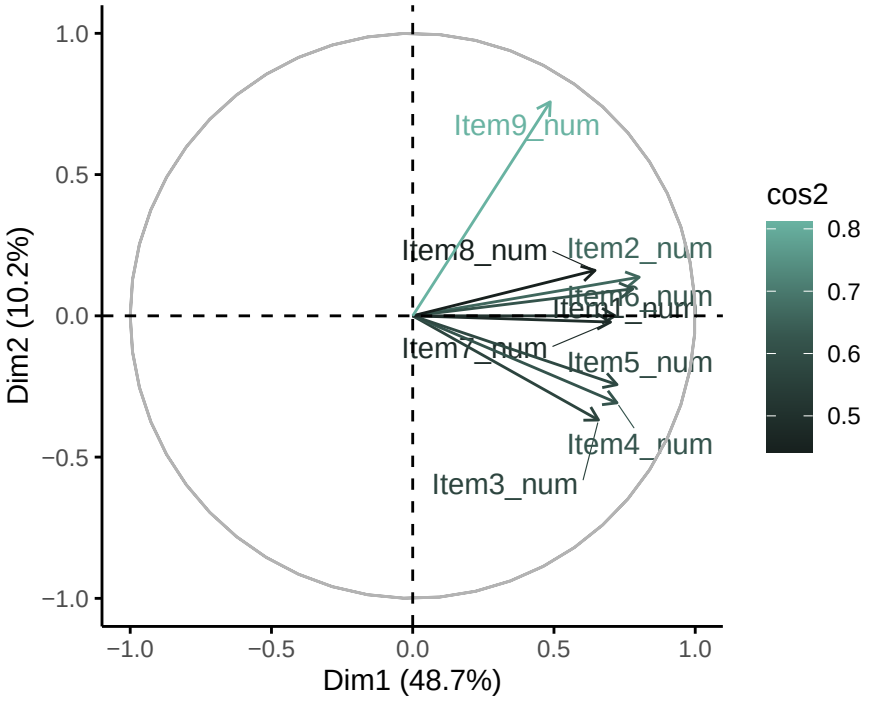

Figure 4. Contribution of variables towards principal component analysis based on cos2 value.

more "difficult" item requires higher latent trait in order to endorse it. For example, Item 9 is the most difficult, because to endorse "Almost every day" requires a latent trait measure of 7.7; which is higher than any other item.

Item Discrimination. If item threshold can be conceptualized as the location along the latent trait measure, then item discrimination is characterized by the slope of the curve. An item is more discriminating if they are better at distinguishing respondents' latent trait measures based on the endorsed response. For example, Item 2, feeling depressed, has the highest discrimination parameter, and from Figure 5, Item 2 has the steepest slope. What this means is that with each step up or down on the latent trait measure, the probability of endorsing a particular response changes a lot more than for other items.

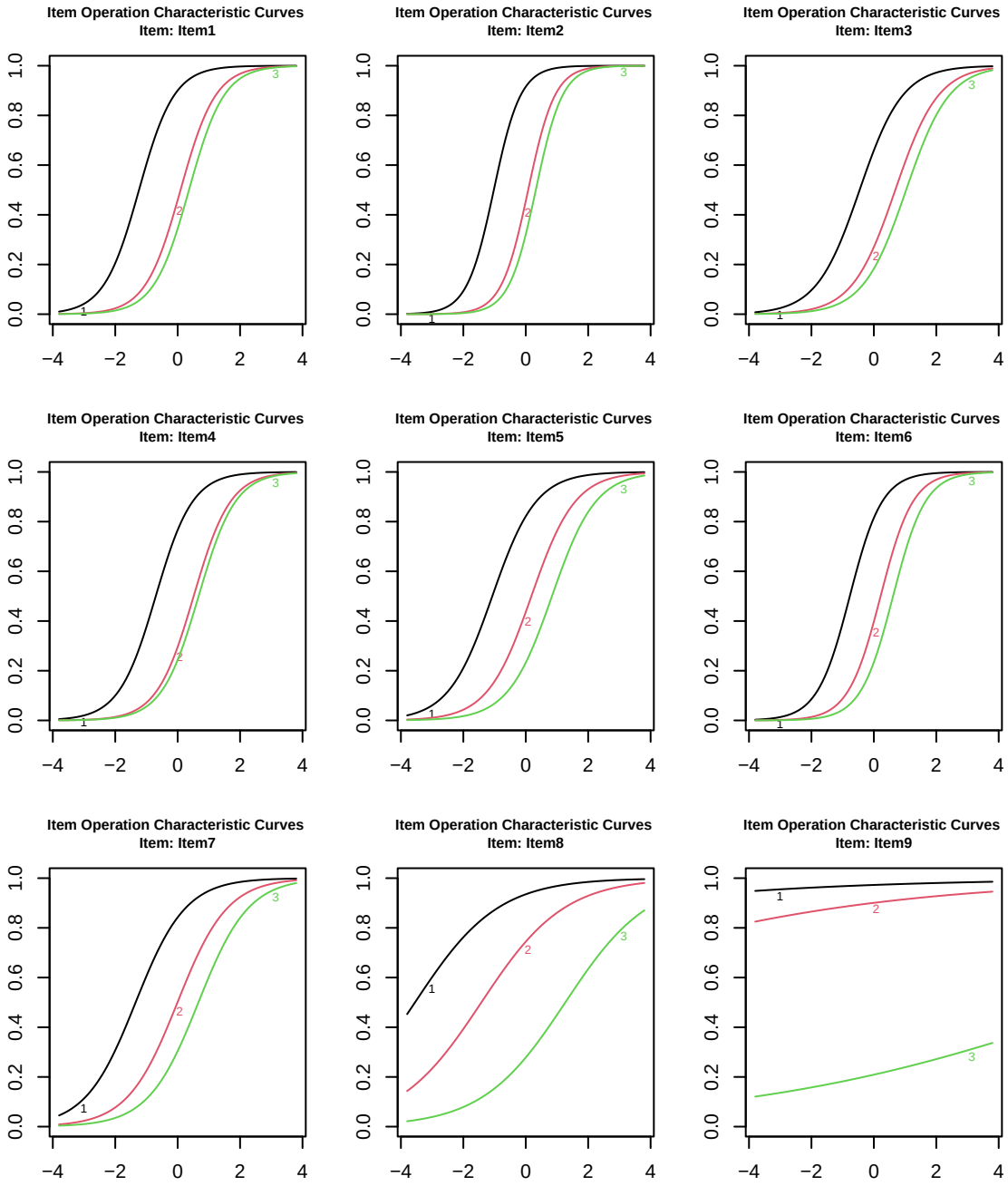

Figure 5. Item Operation Characteristic Curves. Each line corresponds to a category distinction (i.e. 1 corresponds to "Not at all" vs "Several days"; 2 - "Not at all / Several days" vs "More than half the days"; and 3- "Not at all / Several days / More than half the days" vs "Almost every day"). The x-axis corresponds to the latent trait measure, and y-axis depicts the probability of endorsing. Thus, for each line, the item threshold parameter will be at the latent trait measure at 50% probability.

Table 3

Scalability Coefficient H for each PHQ Item

	Coeff H
Item1_num	0.48
Item2_num	0.55
Item3_num	0.46
Item4_num	0.50
Item5_num	0.49
Item6_num	0.53
Item7_num	0.48
Item8_num	0.45
Item9_num	0.39

Item Information. The reliability of a measurement tool within item response theory is conceptualized as information – that is, how much information each item reveals about the latent trait variable. Figure 6 plots the information provided by each item and the entire test. Item 9, wanting to die, and Item 8, moving slowly, provided the least amount of information; whereas Item 2, feeling depressed, provided the greatest amount of information.

Discussion

The typical use case of the PHQ-9 treats responses numerically; summing across all items to derive at a total score. Doing so assumes that each item is weighted similarly towards the final depression score – that is, a respondent who scores "Several days" on both Item 2, feeling depressed, and Item 9, wanting to die, will have their responses be calculated as a 1 for each item. However, analyzing the PHQ-9 scale using item response

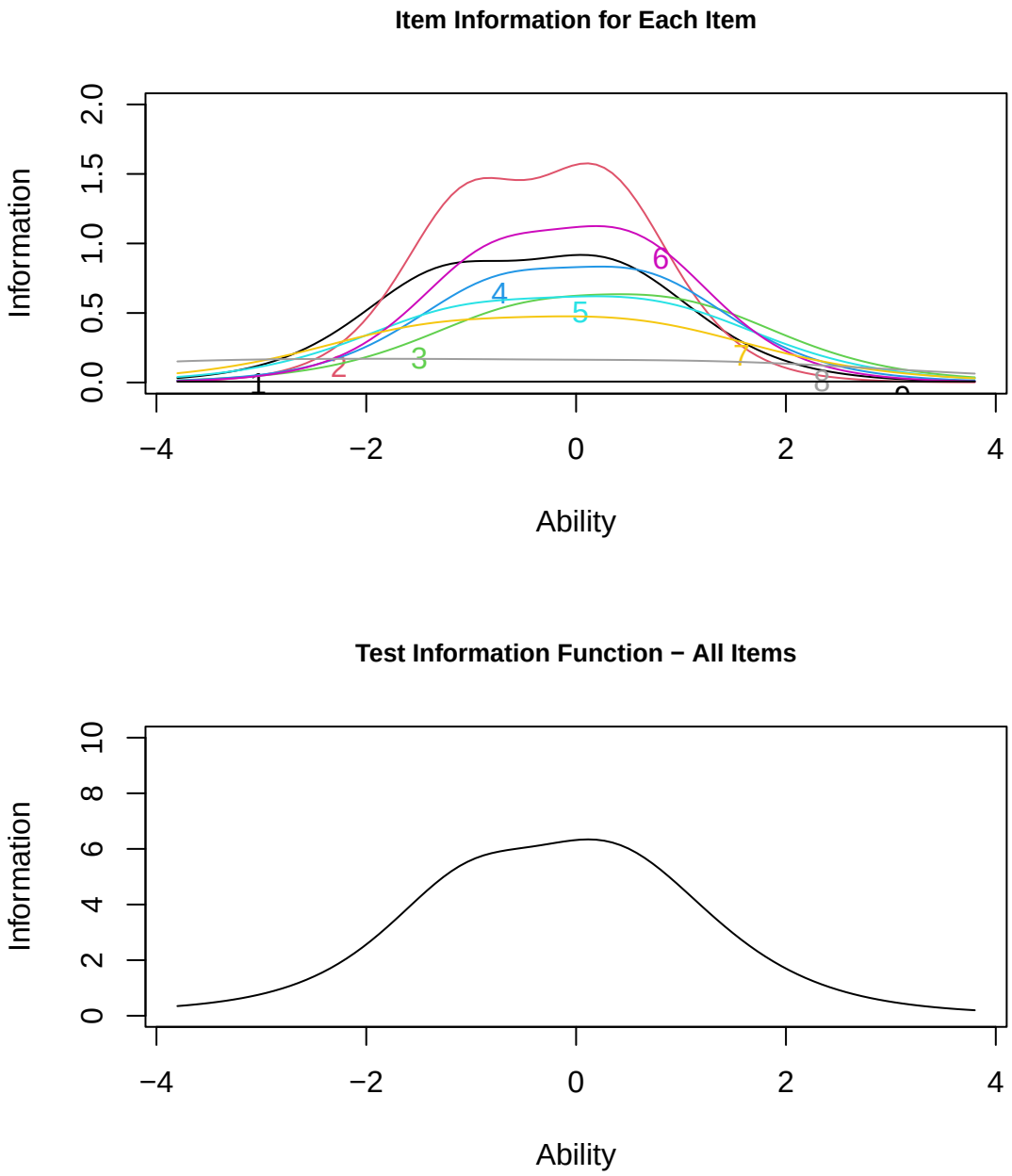

Figure 6. Information Function per Item and Test. Each line corresponds to how much information is provided per item. The second plot of the panel depicts the total information given by the entire test.

Table 4

Item Thresholds and Discrimination Parameters for PHQ-9 Item

	Threshold 1	Threshold 2	Threshold 3	Discrimination
Item1	-1.24	0.09	0.36	1.78
Item2	-1.02	0.08	0.32	2.33
Item3	-0.46	0.69	1.03	1.45
Item4	-0.70	0.51	0.67	1.69
Item5	-1.07	0.18	0.84	1.42
Item6	-0.77	0.22	0.62	1.93
Item7	-1.35	-0.01	0.66	1.25
Item8	-3.55	-1.42	1.27	0.75
Item9	-20.75	-12.81	7.73	0.17

theory shows a more nuanced picture: each item reveals different amounts of information about a respondent's depression, and different items do better at distinguishing among low levels of depression and higher levels of depression. Items 1-6 provides good amounts of information (>0.5) and has good discrimination and thresholds where a respondent's probability of endorsing a response corresponds well to the underlying latent trait measure. Item 9 in particular is did the least well - where most respondents have a high probability of endorsing lower-ranked categories even when depression is high. In the assumptions testing, Item 9 also had the lowest loading in the principle component analysis. An interesting future direction is to explore whether thinking about suicide is a separate dimension from depression, and whether the PHQ scale can be sufficient without including Item 9.

Appendix

Code for Plots

```
# Figure 1: Grid of 3X3 Histograms

PHQ_num <- PHQ[, 2:10] # select only numerical subset
var_name <- colnames(PHQ_num) # get variable names

par(mfrow = c(3, 3), mar=c(3, 3, 3, 3)) # set display grid

# loop through each item to create a histogram and a normal curve
for (i in 1:length(var_name)){
  x = as.numeric(unlist(PHQ_num[var_name[i]][,1]))
  # create histogram
  hist(x, prob = TRUE, breaks = seq(min(x-0.5), max(x+0.5), length.out = 10),
       ylim = c(0,1), main = paste("Item ", i), col = "#69b3a2")
  # create normal distribution
  xfit <- seq(min(x), max(x), length = 2177)
  yfit <- dnorm(xfit, mean=mean(x), sd=sd(x))
  # add curve to plot
  lines(xfit, yfit, col="darkslategray", lwd=2)
}

# Figure 2: Histogram of Sum PHQ-9 Scores

# set x as sum score
x <- as.numeric(unlist(PHQ$SumScore))
```

```r
# create histogram
hist(x, prob = TRUE,
     breaks = seq(min(x), max(x)),
     xlab = "Sum PHQ-9 Scores",
     col = "#69b3a2")
# create normal distribution
xfit <- seq(min(x), max(x), length = 2177)
yfit <- dnorm(xfit, mean=mean(x), sd=sd(x))
# add curve to plot
lines(xfit, yfit, col="darkslategray", lwd=2)

# Figure 3: Scree Plot of Principal Component Analysis

fviz_eig(PHQ_PCA, addlabels = F, ylim = c(0, 50),
         barfill = "#69b3a2",
         barcolor = "darkslategray",
         ggtheme = theme_classic(),
         xlab = "Components",
         main = paste(" "))

# Figure 4: Contributions of Variables to Principle Components

# Color by cos2 values: quality on the factor map
fviz_pca_var(PHQ_PCA, col.var = "cos2",
             gradient.cols = c("#17201e", "#36564e", "#69b3a2"),
             repel = TRUE, # Avoid text overlapping
             title = " ",
```

```
             ggtheme = theme_classic())


# Figure 5: Item Threshold Curves


par(mfrow = c(3, 3), mar=c(3, 2, 3, 2)) # set display grid
plot(mod2, lwd=1, type = "OCCu", ylim = c(0,1), cex.main = 0.8)




# Figure 6: Item And Test Information Function


par(mfrow = c(2, 1)) # set display grid
# plot Item Information Function per item
plot(mod2, type = "IIC",
     main = "Item Information for Each Item",
     ylim = c(0,2))

# plot Test Information Function
plot(mod2, type = "IIC", items=0,
     main = "Test Information Function - All Items",
     ylim = c(0,10))
```

Code for Replicating Analyses

```
# read in data
PHQ <- read_csv("PHQ9_data.csv")
```

```r
# Assumption 1: Unidimensionality
library(FactoMineR)
library(factoextra)
library(corrplot)
# run Principal Component Analysis
PHQ_PCA <- PCA(PHQ_num, scale.unit = TRUE, ncp = 3, graph = FALSE)
PHQ_PCA$eig # display eigenvalues
# correlation plot
var <- get_pca_var(PHQ_PCA)
corrplot(var$cos2, is.corr=FALSE)
# contributions of variables to PC1
fviz_contrib(PHQ_PCA, choice = "var", axes = 1, top = 10)
# contributions of variables to PC2
fviz_contrib(PHQ_PCA, choice = "var", axes = 2, top = 10)


# Assumption 2: Local Independence
library(lavaan)
# testing for one factor, default marker method
model  <- "f  =~ Item1_num + Item2_num + Item3_num + Item4_num +
          Item5_num + Item6_num + Item7_num + Item8_num + Item9_num"
onefactor <- cfa(model, data=phq_num)
summary(onefactor, standardized = TRUE)
# display pathways plot
semPlot::semPaths(onefactor, "std")
fitmeasures(onefactor, c('cfi', 'rmsea', 'rmsea.ci.upper', 'bic'))
# Root Mean Square Error of Approximation (RMSEA) >=0.1
# Comparative fit index (CFI) >=.9
```

```r
# Assumption 3: Monotonicity
library(mokken)
monotonicity.results <- check.monotonicity(PHQ_num)
summary(monotonicity.results) # coeff H >= 0.3 for each item

# Fit Graded Response Model
library(ltm)
PHQ_ordinal <- PHQ[, 12:20] # subset ordinal data
# constrained, i.e. discrimination parameter is held constant
mod1 <- grm(PHQ_ordinal, constrained = TRUE)
mod1
# non constrained model
mod2 <- grm(PHQ_ordinal)
mod2
# compare fit
anova(mod1, mod2)
# extract parameters
coef(mod2)
```

The Patient Health Questionnaire

PATIENT HEALTH QUESTIONNAIRE-9 (PHQ-9)

Over the last 2 weeks, how often have you been bothered by any of the following problems? *(Use "✔" to indicate your answer)*	**Not at all**	**Several days**	**More than half the days**	**Nearly every day**
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself — or that you are a failure or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed? Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
9. Thoughts that you would be better off dead or of hurting yourself in some way	0	1	2	3

FOR OFFICE CODING 0 + _____ + _____ + _____

=Total Score: _____

If you checked off any problems, how difficult have these problems made it for you to do your work, take care of things at home, or get along with other people?

Not difficult at all	**Somewhat difficult**	**Very difficult**	**Extremely difficult**
☐	☐	☐	☐

References

Arrabales, R. (2020). *Perla: A Conversational Agent for Depression Screening in Digital Ecosystems. Design, Implementation and Validation.* https://doi.org/10.48550/ARXIV.2008.12875

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A New Depression Diagnostic and Severity Measure. *Psychiatric Annals*, *32*(9), 509–515. https://doi.org/10.3928/0048-5713-20020901-06

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, *25*(1), 1–18. https://doi.org/10.18637/jss.v025.i01

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Pratap, A., Homiar, A., Waninger, L., Herd, C., Suver, C., Volponi, J., ... Are'an, P. (2022). Real-world behavioral dataset from two fully remote smartphone-based randomized clinical trials for depression. *Scientific Data*, *9*(1), 522. https://doi.org/10.1038/s41597-022-01633-7

R Core Team. (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Revelle, W. (2022). *Psych: Procedures for psychological, psychometric, and personality research*. Evanston, Illinois: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

Rizopoulos, D. (2006). Ltm: An r package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. Retrieved from https://doi.org/10.18637/jss.v017.i05

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Samejima, F. (1997). Graded Response Model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 85–100). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4757-2691-6_5

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, *20*(11), 1–19. Retrieved from https://www.jstatsoft.org/article/view/v020i11

Van der Ark, L. A. (2012). New developments in mokken scale analysis in R. *Journal of Statistical Software*, *48*(5), 1–27. Retrieved from https://www.jstatsoft.org/article/view/v048i05