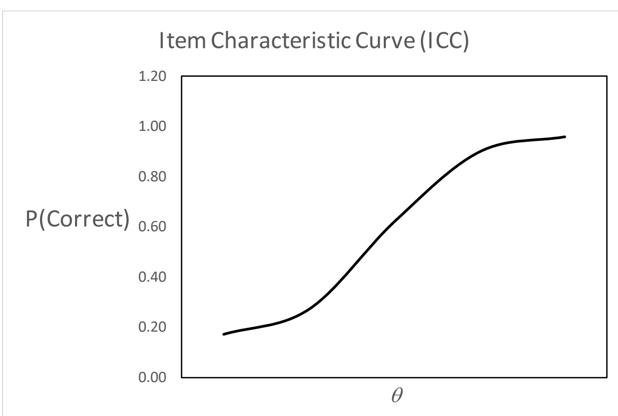
Item Response Models

Item response models are measurement models based on item response theory (IRT: Thurstone, 1925; 1953), and they are used for development and psychometric assessment of measures. Although most commonly employed in an educational testing setting in which aptitude or ability are assessed, item response models can be used to evaluate measures in any area, including attitude measures, behaviors, or traits and individual differences. Item response models generally pertain to measures with multiple binary items assessing some underlying dimension or trait. An example is any aptitude test, such as the GRE in which a set of questions, scored correct or incorrect, are intended to reflect an underlying ability in some area. Although generalizable to ordinal or continuous responses, the typical application of item response models is for a set of binary observed variables. The concept of the underlying ability, trait, or construct is common to psychological and other social science measurement in which the true variable we wish to study is not directly observable but only inferred through multiple particular observations. A single measurement of something is much more fallible than repeated measurements of something. I may measure the length of a room, but I will not likely get the exact same answer each time I measure it. Repeatedly measuring the same thing and then combining the measurements is less subject to error and, thus, more reliable and closer to its true value.[1]

Notation and the IRT Model

Although item response models are useful in many domains, the terminology centers around the ability and correct responses in a testing setting. The unobserved ability which we are intending to assess is designated θ, the Greek letter theta. The essence of the approach is the prediction of whether a particular item on a test is correct or not. This relationship can be represented as a set of logistic regression models in which the ability θ is a predictor of the binary response for each item. If we plot the predicted probabilities that an item is correct as a function of the ability, we would often get the S-shaped curve that is obtained for the relationship between a continuous variable and the probability of a binary response is equal to 1.



This relationship is called the *item characteristic curve* (ICC). The curve should seem quite familiar as the cumulative logistic probability curve, or cdf (see the "Logistic Regression" handout). The regression predicting the item response with the ability trait can be conceptualized as a logistic or a probit regression, however. And like other predictive models in the exponential distribution family, it can be viewed as a generalized linear model with a continuous propensity, y^*, that underlies the observed binary response (see the "Generalized Linear Models" handout). The basic IRT model has two parameters that define this curve, a, the *discrimination parameter*, and b, the *difficulty parameter*. The two parameters are not the familiar regression intercept and slope, however; and, in fact, they are essentially switched. The discrimination parameter a is really the slope, where the steeper the slope the stronger the

[1] Multiple measures are more accurate, assuming we really are on target for what we are trying to measure (i.e., it is a valid measure). In the case of length, it is hard to imagine something other than length being assessed with a ruler, but we may be less certain that the verbal section of the GRE is actually measuring underlying verbal abilities rather than something else, such as reading speed.

relationship between the ability and a correct response, giving an indication of how well a correct response discriminates on the ability. The difficulty parameter b, represents the point at which half (p = .5) of the respondents get the item correct. The intercept is equal to $-ab$, so, if α is the intercept from the right-hand side of the equation in the usual generalized linear model, then the difficulty parameter is $b = \alpha/\text{-}a$.

One of the more common IRT models is the *normal ogive model*, which is essentially just the probit model (ogive here is another word for the cumulative normal distribution). Because the model as described above has two parameters, it is called the *two-parameter logistic model* (2PL) or the *two-parameter normal ogive model* (2PN). Expressed in terms of the unobserved continuous response, denoted by z in this context, the 2PN model is written as

$$z = a(\theta - b)$$

If there is no difference between the ability θ and the difficulty b, such that $\theta - b = 0$ and $\theta = b$, then the z-score is 0 and the probability that the item is correct is .5. In this form, it does not clearly resemble the generalized linear regression model, but knowing that $\alpha = -ab$, and using a little algebra, we can see the connection much more clearly.

$$\begin{aligned} z &= a\theta - ab \\ y^* &= \beta X - \alpha \\ &= -\alpha + \beta X \end{aligned}$$

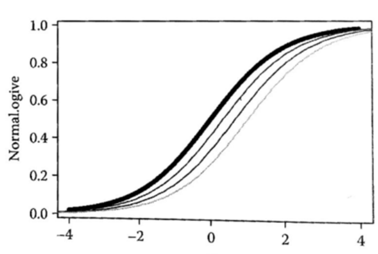
So, it really is a generalized linear model, except that the intercept is re-expressed to provide the difficulty parameter as a function of both the slope and the intercept, the predictor is the unobserved ability θ, and the intercept has a negative sign.

One can compute the predicted probability of a correct response (that Y = 1) for a given ability value using the logistic cdf equation, where the usual regression equation is replaced by the IRT equation:

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

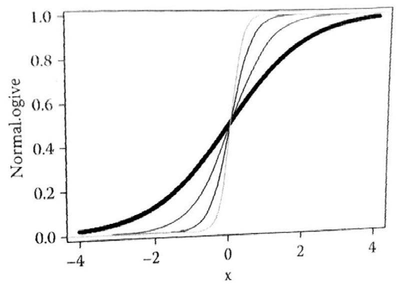
Values are sometimes modified using the constant multiplier 1.7 which approximately converts the logistic scaling to a normal scaling (for reference to the normal ogive).

The ICC for several items can be plotted together to illustrate the interpretation of two parameters. In the curve below, the items vary on their difficulty (i.e., the b parameters differ) but they have the same relationship to the ability, so have the same level of discrimination (i.e., the a parameters are the same). As the darker line has the higher probability of correct response overall (i.e., higher "difficulty").



Raykov and Marcoulides (2011, p. 280)

The second ICC shows that the items vary in how they discriminate (i.e. have different a parameters), where the dark line represents poorer discrimination of the ability trait.



Raykov and Marcoulides (2011, p. 279)

Information Function

A quantity related to the a and b parameters is *information*. The information function is connected to the precision or standard error for a particular item (or can be used for the total scale, in which it is the total of the information of all of the items). For binary variable in the two-parameter model, the information value is a function of the first derivative (indicated by ′) of the probability of correct response relative to the variability estimate (DeMars, 2018):[2]

$$I_i(\theta) = \frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)\left[1 - P_i(\theta)\right]'}$$

The information of a particular item will be greater for moderate values of b (moderate difficult) and high values of a (high discrimination) and is ideally highest near the ability level for the test taker. Curves of the information for individual items (or the scale) are often examined, with more peaked curves near the ability preferable because they provide the most information about the ability.

Variations on the Common Two-Parameter Model

IRT models can be estimated as logistic or probit models. A special case of the IRT model discussed above assumes that the discrimination parameter, a, for all items is equal but allows items to differ in the difficulty parameter b (as depicted in the first figure above), which is a one-parameter (or 1PL) model. When this is estimated as a logistic model, it is known as the *Rasch model* (Rasch, 1960). The three-parameter model (3PL or 3PN) adds a *guessing parameter* that takes into account the chance of getting the item correct, adding a low asymptote or minimum probability at which point those low on ability would get the item correct. The guessing parameter, c, should be equal to or better than chance if the test taker can guess the item correctly. When the alternative on multiple choice question (the destractors) are particularly good, a person with low ability should be less likely than chance to correctly guess the answer. The three-parameter model may not be relevant or needed in many circumstances in which guess in not an issue (e.g., personality test). IRT models can be extended to ordinal (called graded response, Samejima, 1969) or multiple category responses is using the other generalized linear regression models we have discussed.

[2] The term “information” is derived from the Fisher information matrix, which is commonly used in computation of asymptotic standard error estimates with maximum likelihood.

Software

There is no IRT procedure in SPSS. In R, IRT models can be estimated with `mirt`, `TAM`, `ltm`, or `eRm` (for Rasch modeling approach) among many other packages. In SAS, PROC IRT procedure (and IRTFIT-RESAMPLE, DRAWICC, and IRGEN macros) estimates IRT models. There also are some well-known standalone packages like MULTILOG (Thissen, Chen, & Bock, 2002; and former BILOG) and PARSCALE (Muraki & Bock, 1991). The IRT model involves an unobserved trait, so can be conceptualized and estimated as a latent variable model using structural equation modeling software that is capable of estimation with binary observed variables (e.g., Mplus, lavaan package in R, Lisrel).

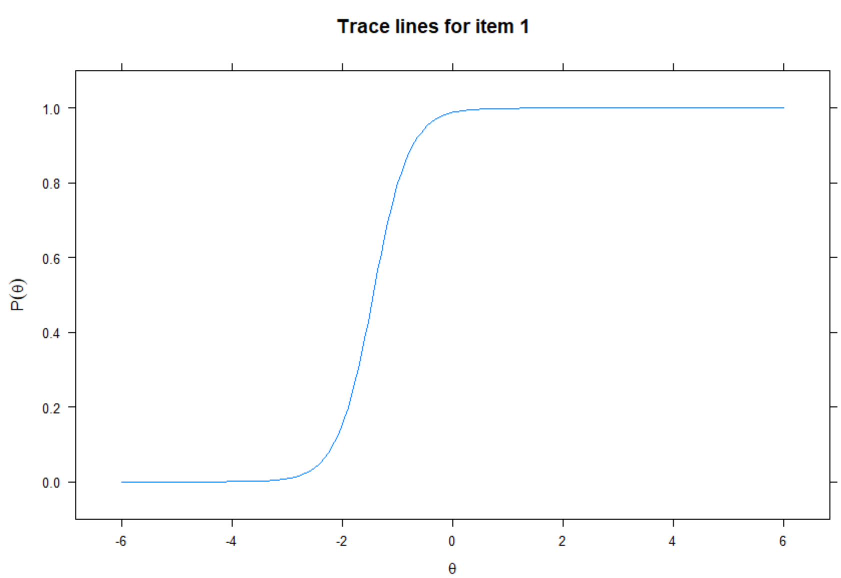
Example

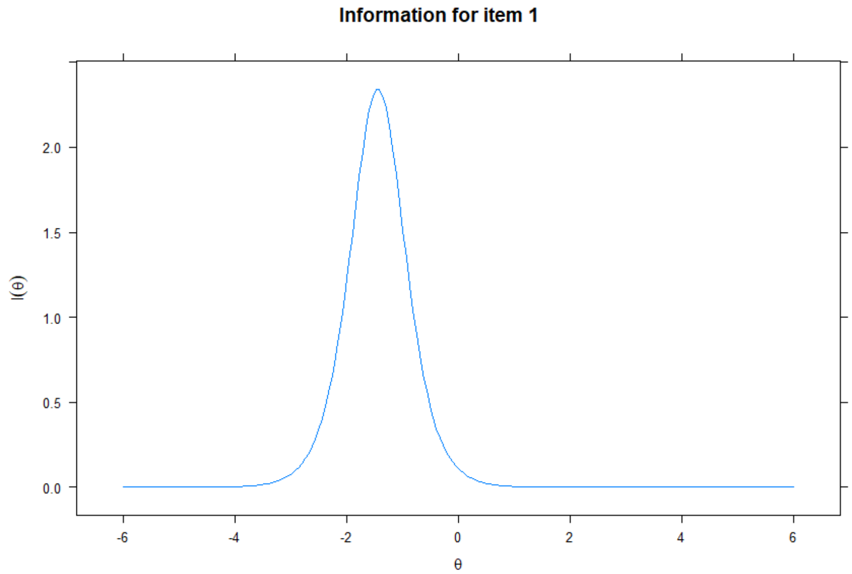
Data for this example come from the National Caregiver Health Effects study (Schulz et al., 1997) using a subset of participants and items from the Mini-Mental State Exam (MMSE), a screen for cognitive dementia. I test a two-parameter IRT model with thirteen binary items (incorrect = 0, correct = 1) from questions about the date, city, or recall of a word. I use the `mirt` package in R (Chalmers, 2012) and PROC IRT in SAS. Some parts of the output or some plots are omitted to save space.

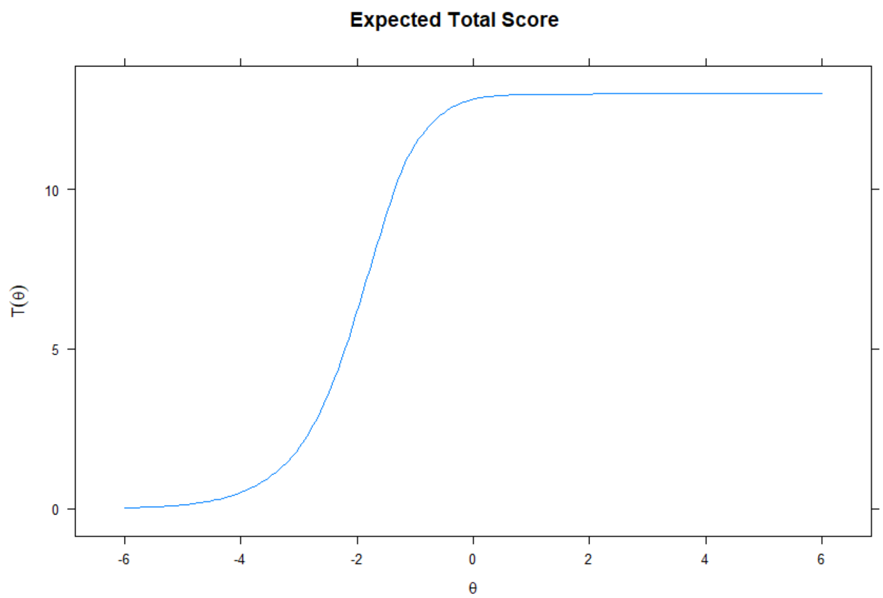
R

```
> #use a listwise deletion routine to eliminate missing data
> d <- d[complete.cases(d), ]
>
> library("mirt")
> irtmod <- mirt(data = d, model = 1, itemtype = "2PL")
Iteration: 151, Log-Lik: -648.571, Max-Change: 0.00010
> summary(irtmod)
             F1    h2
year33    0.874 0.764
season33 0.780 0.608
date33    0.892 0.795
day33     0.902 0.814
month33   0.969 0.939
state33   0.899 0.809
county33 0.869 0.755
city33    0.956 0.913
floor33   0.864 0.747
addrss33 0.889 0.790
apple133 0.696 0.484
table133 0.668 0.446
penny133 0.705 0.498

SS loadings:  9.361
Proportion Var:  0.72

Factor correlations:

   F1
F1  1
> print(irtmod)

Call:
mirt(data = d, model = 1, itemtype = "2PL")

Full-information item factor analysis with 1 factor(s).
Converged within 0.0001 tolerance after 151 EM iterations.
mirt version: 1.33.2
M-step optimizer: BFGS
EM acceleration: Ramsay
Number of rectangular quadrature: 61
Latent density type: Gaussian

Log-likelihood = -648.571
Estimated parameters: 26
AIC = 1349.142; AICc = 1354
BIC = 1446.791; SABIC = 1364.326
G2 (8165) = 212.14, p = 1
RMSEA = 0, CFI = NaN, TLI = NaN

> coef(irtmod, IRTpars = TRUE)
$year33
        a      b g u
par 3.064 -1.444 0 1

$season33
        a      b g u
par 2.118 -2.138 0 1
```

```
$date33
        a      b g u
par 3.352 -0.671 0 1

$day33
        a      b g u
par 3.556 -1.397 0 1

$month33
       a      b g u
par 6.65 -1.419 0 1

$state33
        a      b g u
par 3.502 -2.421 0 1

$county33
        a      b g u
par 2.984 -2.009 0 1

$city33
        a      b g u
par 5.518 -1.956 0 1

$floor33
        a      b g u
par 2.924 -2.539 0 1

$addrss33
        a      b g u
par 3.304 -1.842 0 1

$apple133
        a      b g u
par 1.649 -3.406 0 1

$table133
        a      b g u
par 1.527 -2.629 0 1

$penny133
        a      b g u
par 1.694 -2.535 0 1

$GroupPars
    MEAN_1 COV_11
par      0      1

> #produces plot for the expected score of the total scale
> plot(irtmod)
> #item plots allowed one at a time using item number, trace is ICC and info is information
> itemplot(irtmod,1,type="trace")
> itemplot(irtmod,1,type="info")
```







SAS

```
ods graphics on;
proc irt data=one plots=(scree icc iic tic);
var Year33 Season33 Date33 Day33 Month33
State33 County33 City33 Floor33 Addrss33
Apple133 Table133 Penny133;
run;
```

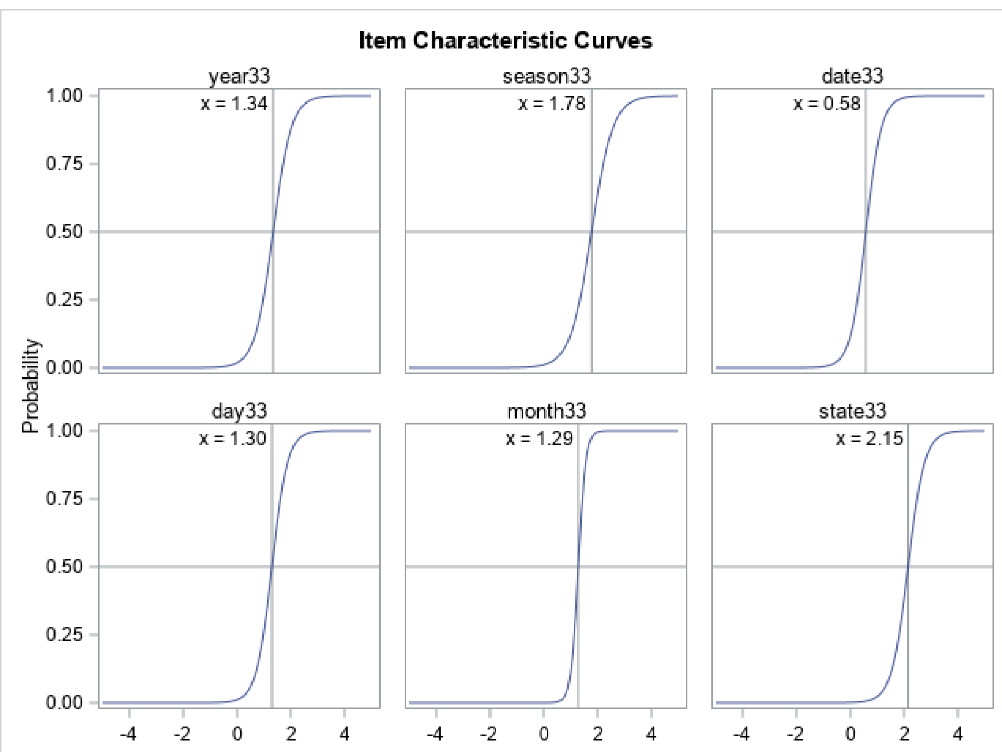
The IRT Procedure

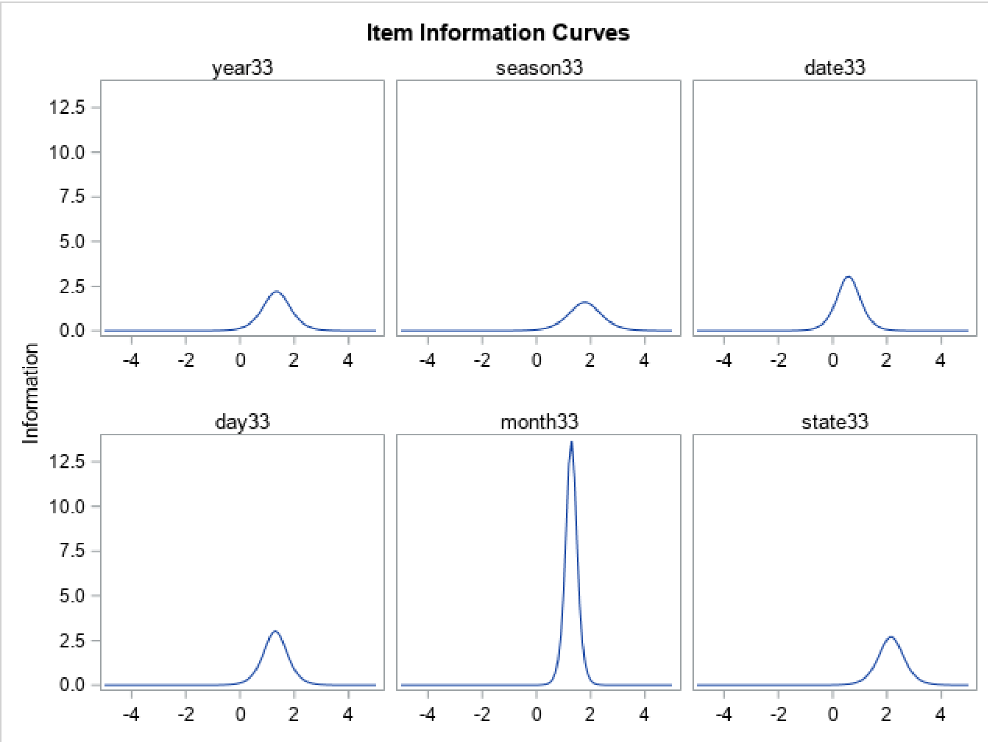
Model Fit Statistics

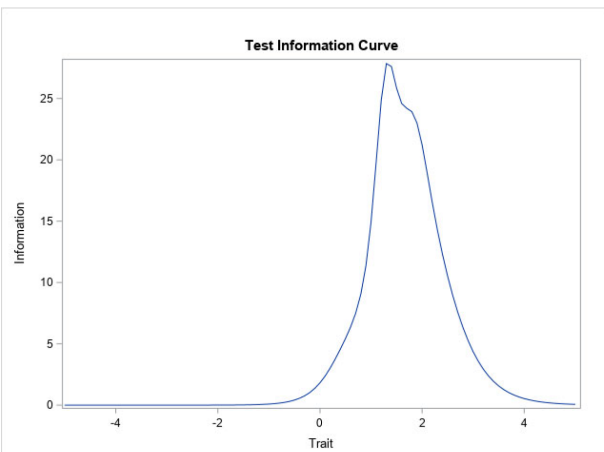
Log Likelihood	-761.7332367
AIC (Smaller is Better)	1575.4664735
BIC (Smaller is Better)	1674.3215512
LR Chi-Square	218.6937473
LR Chi-Square DF	8165

Item Parameter Estimates

Item	Label	Parameter	Estimate	Standard Error	Pr > |t|
year33	WHAT IS THE YEAR	Difficulty	1.34261	0.12828	<.0001
		Slope	2.97276	0.58199	<.0001
season33	WHAT IS THE SEASON	Difficulty	1.78323	0.19314	<.0001
		Slope	2.53367	0.54498	<.0001
date33	WHAT IS THE DATE	Difficulty	0.58212	0.08609	<.0001
		Slope	3.49689	0.77757	<.0001
day33	WHAT DAY OF WEEK	Difficulty	1.29658	0.11661	<.0001
		Slope	3.48363	0.71139	<.0001
month33	WHAT IS THE MONTH	Difficulty	1.28670	0.10043	<.0001
		Slope	7.39614	2.98498	0.0066
state33	WHAT STATE ARE WE IN	Difficulty	2.14945	0.26423	<.0001
		Slope	3.29217	1.05228	0.0009
county33	WHAT COUNTY ARE WE IN	Difficulty	1.79550	0.18786	<.0001
		Slope	3.23502	0.81934	<.0001
city33	WHAT CITY ARE WE IN	Difficulty	1.84201	0.16811	<.0001
		Slope	5.36665	1.91865	0.0026
floor33	WHAT FLOOR (BUILDING) ARE WE ON	Difficulty	2.31928	0.30445	<.0001
		Slope	2.95911	0.91823	0.0006
addrss33	WHAT IS THIS ADDRESS	Difficulty	1.68243	0.16261	<.0001
		Slope	3.46826	0.84813	<.0001
apple133	RECALL APPLE (1)	Difficulty	2.56288	0.38103	<.0001
		Slope	2.35696	0.69880	0.0004
table133	RECALL TABLE (1)	Difficulty	2.24028	0.32142	<.0001
		Slope	1.75430	0.40465	<.0001
penny133	RECALL PENNY (1)	Difficulty	2.18202	0.29559	<.0001
		Slope	1.91920	0.44124	<.0001







References and Further Reading

Chalmers, R. P. (2012). *mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software, 48*(6), 1-29.
DeMars, C. (2010). *Item response theory.* Oxford University Press.
DeMars, C. (2012). A Comparison of Limited-Information and Full-Information Methods in Mplus for Estimating Item Response Theory Parameters for Nonnormal Populations. *Structural Equation Modeling, 19*, 610-632.
DeMars, C. (2018). Item information function. In B.B. Frey (Ed.). (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation.* Sage Publications.
Kamata, A., & Bauer, D.J. (2008). A Note on the Relation Between Factor Analytic and Item Response Theory Models. *Structural Equation Modeling, 15*,136–153.
Muraki, E., & Bock, R. D. (1991). *PARSCALE: Parameter scaling of rating data [Computer program].* Chicago, IL: Scientific Software, I
Rasch G (1960). *Probabilistic Models for some Intelligence and Attainment Tests.* Danish Institute for Educational Research, Copenhagen.
Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York: Routledge.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 17*(4), 2. doi:10.1002/j.2333-8504.1968.tb00153.x
Schulz, R., Newsom, J., Mittelmark, M., Burton, L., Hirsch, C., & Jackson, S. (1997). Health effects of caregiving: the caregiver health effects study: an ancillary study of the Cardiovascular Health Study. *Annals of Behavioral Medicine, 19*(2), 110-116.
Thissen, D., Chen, W.-H., & Bock, R. D. (2002). MULTILOG (Version 7.03) [Computer software]. Lincolnwood, IL: Scientific Software International.
Thurstone, L. L. (1925). A method of scaling psychological and educational tests. Journal of Educational Psychology, 16, 433-445.
Thurstone, L. L. (1953). *S.R.A. Primary Abilities for ages 5 to 7.* Chicago: Science Research Cooperation