Extensions of Item Response Models

Fit

Maximum likelihood estimation is generally used for fitting item response theory models. Theoretically, the fit of the model reflects the discrepancy between the expected and observed values in the items based on a single dimension of ability but the lack of fit may be due to a variety of factors including assumption violations. Deviance (-2 loglikelihood) and likelihood ratio chi-square values are produced in IRT software, but these values are not terribly informative by themselves as there is no standard by which to judge whether a model has adequate fit without comparison to alternatives. Model fit can be used to compare the appropriateness of one-, two-, and three-parameter models, however. The Hausman tests compares the usual marginal maximum likelihood fit to the fit obtained with a limited information estimation and, although not widely available, appears to be a good assessment of global fit if the samples size is 1000 or more (Ranger & Much, 2020). Another metric that is sometimes mentioned is the coefficient of reproducibility, which is a function of the total number of errors, $C_R = 1 - (total\ errors/total\ repsonses)$. Some authors (e.g., Thorpe & Favia, 2012; Kline, 2005) give .85 or .90 as acceptable cutoffs for the coefficient of reproducibility, but this metric is not necessarily an indication of whether the test overall is a reliable one or all of the items are good items.

Multiple Dimensionality

Multidimensional IRT (sometimes MIRT), in which the underlying ability has two or more subdomains (e.g., verbal and math ability) is also possible (McDonald, 1997; Rekase, 2009). Most often in practice, the investigation of the number of dimensions has involved principal components analysis (PCA) or factor analysis. IRT procedures often print eigenvalues (which pertain to the number of possible underlying dimensions) and loadings (the linear relationship between the ability and the item) from a principal components analysis. Scree plots of the magnitude of the eigenvalues, typically used to select the number of dimensions in principal components analysis are commonly generated by IRT software. When there are theoretical or empirical reasons to believe there are multiple dimensions, a common approach has been to conduct subsequent IRT analysis separately for the different constructs (Turk et al., 2006). Increasingly, either confirmatory factor analysis or more flexible IRT software has been used to conduct IRT analyses when multiple dimensions are present (e.g., Cai, 2010; Rindskopf & Rose, 1988).

Differential Item Function

Differential item functioning (DIF) refers to differences in the IRT parameters across groups,[1] either differences in the difficulty (uniform DIF) or the discrimination parameter (non-uniform DIF). For example, in a two-parameter model, either the difficulty b, or the discrimination parameter, a, or both, could differ across groups. If bias exists for a particular item, one typically expects that the relationship between the ability and the probability of a correct response on the item (i.e., the item is a better reflection of ability in one group than another). The converse—that the presence of DIF indicates bias—is not necessarily true. DIF however is a preferable approach to investigating bias than examining proportion or mean differences on items or the scale. Differences across groups in the overall probability of a correct response on an item (base rate of a correct response) is generally not considered to be evidence of bias, because it has not been ruled that the difference across groups is not reflective of overall differences in ability between the groups. Graphical methods (e.g., plotting group ICCs next to one another) or statistical tests can be conducted to investigate DIF. Statistical tests can be conducted with a Cochran-Mantel-Haenszel chi-square, logistic regression, or a likelihood ratio test using equality constraints. The logistic model approach requires regressing the item on both the ability (total score) and the group variable to show that the effect of the group predict the item over and above the ability. The logistic approach is best when the total score does not include the item of interest.

[1] The variable that is the potential source of bias does not necessarily have to be a grouping variable. It could be continuous.

Graded Response Models

The graded response model (sometimes GRM) is an extension of IRT models for ordinal items (Muraki, 1990; Samejima,1969). Graded response models involve a regression of each ordinal item on the ability construct, with the familiar slope and threshold estimates used with the y^* ordinal logistic or ordinal probit (in IRT, the normal ogive model) interpretations. Instead of item characteristic curve (ICC), however, graded response models use the term category response curve (CRC). Predicted probability is from one ordinal category to the next, assuming equal odds just as in the ordinal logistic or probit models. In terms of the logistic version of the model, we can compute the probability of increment on the ordinal scale for an item, j, using the logistic cdf equation subtracting the probability of response of one category from the probability of response, K, from the next lower category, K - 1.[2]

$$P\left(Y_{ij} = k \mid \theta_j\right) = \frac{1}{1+e^{-a_j\left(\theta - b_{jK-1}\right)}} - \frac{1}{1+e^{-a_j\left(\theta - b_{jK}\right)}}$$

Computation of the probability from the normal ogive requires the normal cdf translation that was discussed earlier with probit models. The CRCs for each category may be plotted together for each item to examine the difficulty and discrimination at each level of the item response. Note that the discrimination (slope or a) parameter is usually assumed to the same across ordinal levels of the item variable. The difficulty parameter, b, can differ. Item information (or either item or total score) can also be computed, although the computation is slightly more complicated than for the binary case and, when plotted, will have multiple peaks with the maximum number of peaks corresponding to number of ordinal response categories.

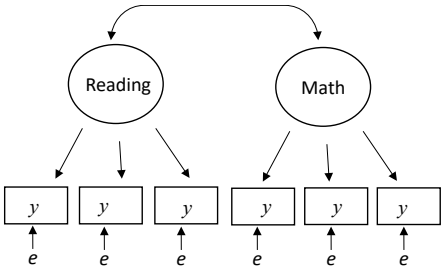
In SAS you can use the `/resfunc=gr` option on the model line in PROC IRT and, in R, you can use `itemtype = 'graded'` in the `mirt` package.

Connection to Factor Analysis

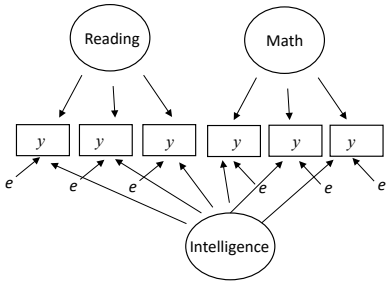
Although IRT researchers usually distinguish item response theory from classical test theory, there is a clear connection (equivalence really) between IRT models and factor analysis models (e.g., Reise & Widaman, 1999). Both involve the estimation of the association between an unobserved trait and responses on a particular item. Factor analysis can be conducted with binary or ordinal variables, although usually these models are conducted with confirmatory factor analysis (structural equation modeling) programs, such as Mplus, R lavaan, or LISREL. The most common approaches to these models when the items are binary are the marginal maximum likelihood approach (Christoffersson, 1975) which usually gives logistic estimates (loadings), or a weighted least squares approach (Muthén, du Toit, & Spisic, 1997), which gives probit estimates. The loadings and intercepts (thresholds) from these programs are not in the IRT formulation for a and b parameters but are instead true regression slopes and intercepts (see the "Item Response Theory" handout for a discussion). The estimates can be converted to the IRT parameter values, whether in logistic or probit form, however (see Kamata & Bauer, 2008 for details). Otherwise the models are equivalent.

There are several ways to specify multiple dimensions/factors with each approach. Immekus, Snyder, and Ralston (2019) give a good overview. Below, on the left, is the typical two-factor confirmatory factor analysis model (here, assume binary indicators). The bifactor is a common approach for IRT models (e.g., Reise, 2012), which would be specified as shown below on the right. The illustration depicts two dimensions (reading and math) that predict item responses over and above the overall ability (intelligence).

[2] I am following the equations in our reading in that the $1/1+e$ form of the question is use (as opposed to the $e/1+e$ form) which is for the probability at or above and therefore the probability for k is subtracted from $k-1$.



Common two-factor CFA



IRT Bi-Factor Model

Immerkus and colleagues also discus a second-order factor (or two-tiered) model that can be specified in CFA (e.g., Rindskopf & Rose, 1988) or estimated with some IRT programs (e.g., Cai, 2010).

Examples

The analyses below investigate DIF males and females on five items from a verbal ability test from the International Cognitive Ability Resource (ICAR) repository.[3] Each of the five items are binary correct or incorrect responses. In R `mirt`, I compared models with and without all of the slopes constrained. The `anova()` function create a likelihood ratio test to compare their fits. In SAS, I compared just one item across groups. I've omitted some of the output to save space.

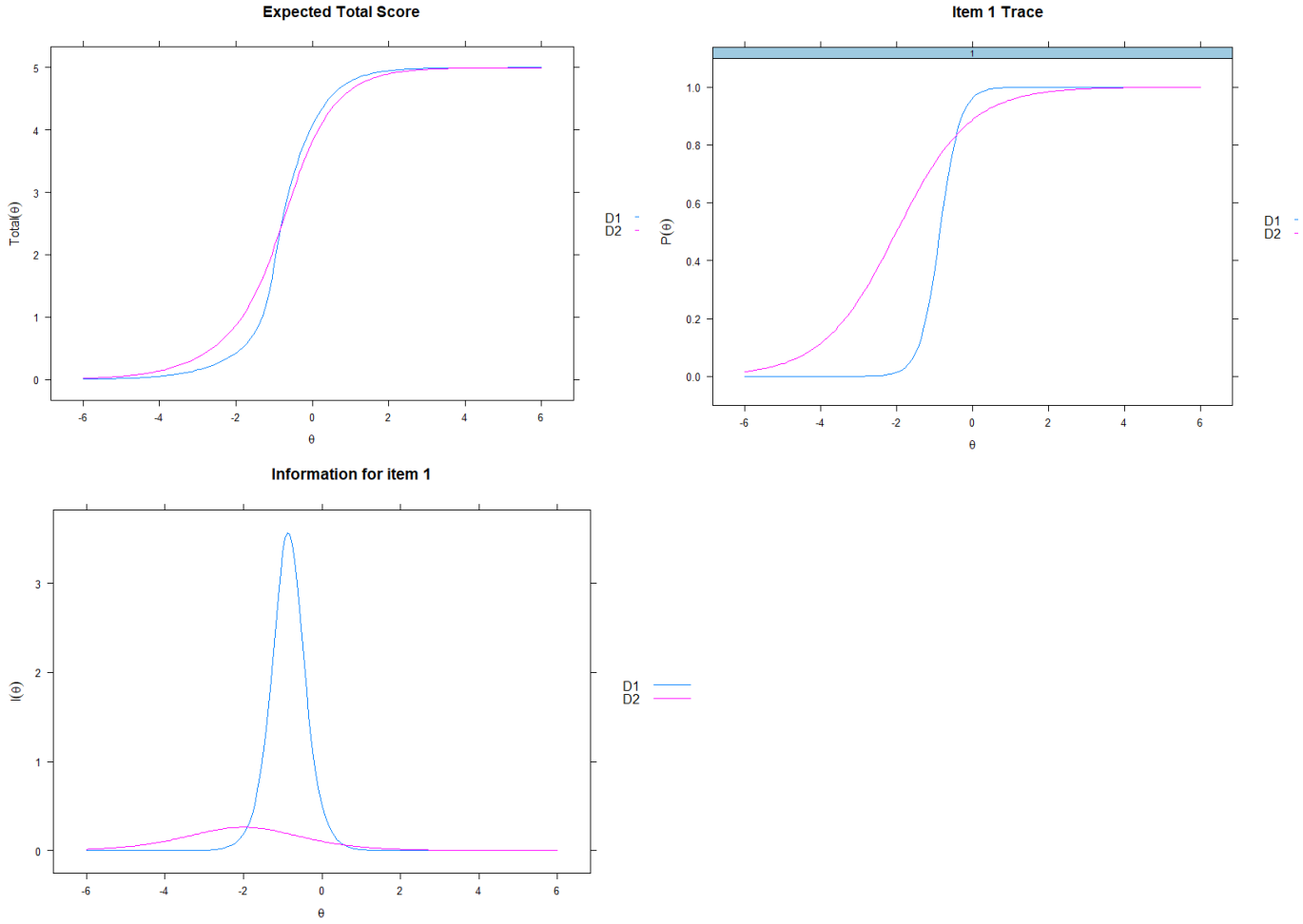
R

```
> # first sort cases by the group variable, then create a new group based on the Ns from each half
> d <- d[order(d$sex),]
> group <- c(rep('D1', 65), rep('D2', 133))
> #use lessR to subset only the numeric item variables
> library(lessR)
> d2 = Subset(columns=c(v2, v4, v5, v6, v8))

> library("mirt")

> irtmod1 <- multipleGroup(d2, model = 1, group = group)
> irtmod2 <- multipleGroup(d2, model = 1, group  = group, invariance = c('slopes'))

> anova(irtmod1,irtmod2)

Model 1: multipleGroup(data = d2, model = 1, group = group, invariance = c("slopes"))
Model 2: multipleGroup(data = d2, model = 1, group = group)

> coef(irtmod2, IRTpars = TRUE)
$D1
$v2
        a      b g u
par 1.678 -1.157 0 1

$v4
        a      b g u
par 2.202 -1.271 0 1

$v5
        a      b g u
par 1.621 -0.182 0 1

$v6
        a     b g u
par 1.738 -1.22 0 1

$v8
        a      b g u
par 1.537 -0.369 0 1

$GroupPars
    MEAN_1 COV_11
par      0      1


$D2
```

[3] Condon, D. M., & Revelle, W. (2016). Selected ICAR Data from the SAPA-Project: Development and Initial Validation of a Public-Domain Measure. Journal of Open Psychology Data, 4(1), e1.DOI: http://doi.org /10.5334/jopd.25

```
$v2
        a      b g u
par 1.678 -1.463 0 1

$v4
        a      b g u
par 2.202 -0.478 0 1

$v5
        a      b g u
par 1.621 -0.402 0 1

$v6
        a      b g u
par 1.738 -0.912 0 1

$v8
        a      b g u
par 1.537 -0.529 0 1

$GroupPars
    MEAN_1 COV_11
par      0      1
```

```
> #produces plot for the expected score of the total scale
> plot(irtmod1)
> #item plots allowed one at a time using item number, trace is ICC and info is information
> itemplot(irtmod1,1,type="trace")
> itemplot(irtmod1,1,type="info")
```



SAS

SAS PROC IRT wants the grouping variable to be coded 1 and 2, so I compute a new variable first. The equality option allows the user to impose constraints across groups on a single item (here, I constrained v8), or multiple items, and either the difficulty parameter [intercept] or discrimination parameter [slope] or

both (same statements but separate by a comma). Some output has been omitted to save space (the item parameters are from the model without constraints).

```
data two; set one;
if sex=1 then sexgrp=2;
if sex=0 then sexgrp=1;
run;

proc freq data=two;
run;

*see Zhang 2015 SAS white paper on DIF;

ods graphics on;
proc irt data=two plots=(scree icc iic tic);
var v2 v4 v5 v6 v8;
group sexgrp;
run;

proc irt data=two plots=(scree icc iic tic);
var v2 v4 v5 v6 v8;
group sexgrp;
equality v8 /parm=[slope] between_gp=[1 2];
run;
```

Unconstrained Model Parameters

```
The IRT Procedure
                                Model Fit Statistics

                          Log Likelihood             -513.2356301
                          AIC (Smaller is Better)    1066.4712603
                          BIC (Smaller is Better)    1132.2366009
                          LR Chi-Square              21.993395191
                          LR Chi-Square DF                     43

                               Item Parameter Estimates
                                       sexgrp = 1
                                                                 Standard
Item    Label    Parameter       Estimate         Error    Pr > |t|

v2      v2       Difficulty      -0.86040       0.21562      <.0001
                 Slope            3.74945       1.86719      0.0223
v4      v4       Difficulty      -0.97362       0.20270      <.0001
                 Slope            7.18842       8.59941      0.2016
v5      v5       Difficulty      -0.19434       0.23204      0.2011
                 Slope            1.58286       0.62009      0.0053
v6      v6       Difficulty      -1.46738       0.52646      0.0027
                 Slope            1.18601       0.51026      0.0101
v8      v8       Difficulty      -0.29773       0.19315      0.0616
                 Slope            2.59787       1.07634      0.0079

                                    sexgrp = 2
                                                                 Standard
Item    Label    Parameter       Estimate         Error    Pr > |t|

v2      v2       Difficulty      -2.00256       0.65656      0.0011
                 Slope            1.02889       0.42797      0.0081
v4      v4       Difficulty      -0.51420       0.16916      0.0012
                 Slope            1.86071       0.60700      0.0011
v5      v5       Difficulty      -0.44692       0.19674      0.0116
                 Slope            1.33482       0.43934      0.0012
v6      v6       Difficulty      -0.81359       0.18102      <.0001
                 Slope            2.40359       0.92748      0.0048
v8      v8       Difficulty      -0.58986       0.21482      0.0030
                 Slope            1.27609       0.39315      0.0006
```

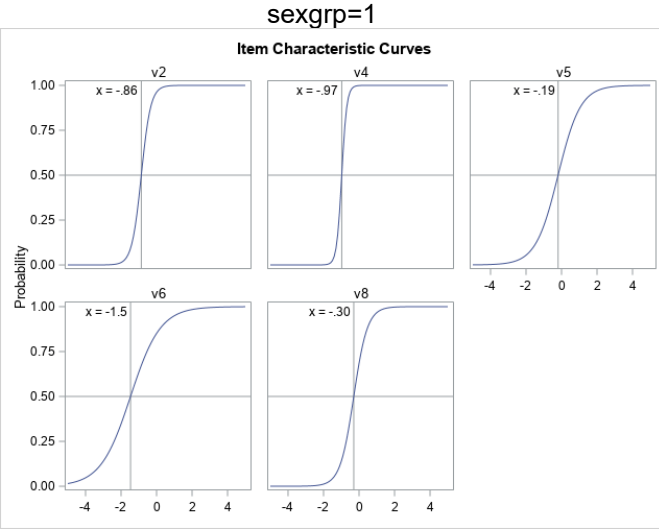
Constrained Model Fit

```
The IRT Procedure

                    Model Fit Statistics

Log Likelihood              -514.2091419
AIC (Smaller is Better)     1066.4182837
BIC (Smaller is Better)     1128.8953573
LR Chi-Square               22.853783658
LR Chi-Square DF                      44
```

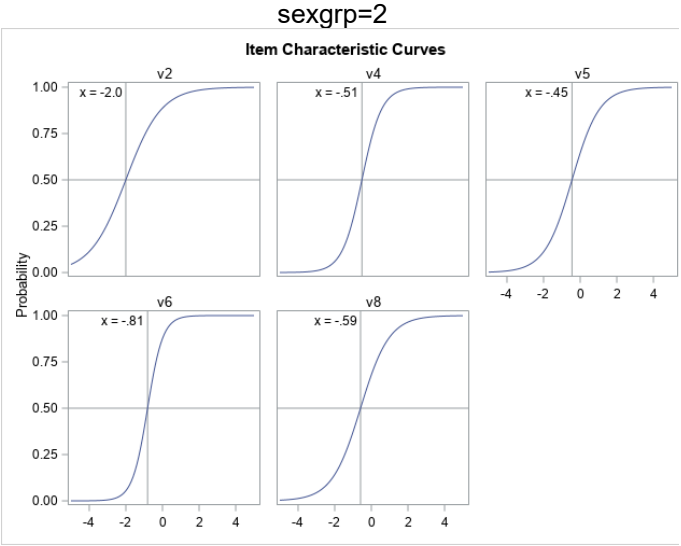
Difference between the two chi-squares is a LR test of (non-uniform) slope DIF, which I computed by hand: 22.853783658 - 21.993395191 = 0.860388467 with 44- 43 = 1 df, which is not significant.

```
Obtain separate ICC plots
ods graphics on;
proc irt data=two plots=icc;
var v2 v4 v5 v6 v8;
by sexgrp;
run;
```

sexgrp=1



sexgrp=2



References and Further Reading

Bauer D.J. (2017) A more general model for testing measurement invariance and differential item functioning. *Psychological Methods, 22*(3):507.

Cai, L. (2010). A two-tiered full-information item factor analysis model with applications. *Psychometrika 75,* 581–612. doi: 10.1007/s11336-010-9178-0

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40*, 5–32.

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. doi:10.1080/10705511.2012.713272

Immekus, J. C., Snyder, K. E., & Ralston, P. A. (2019, May). Multidimensional item response theory for factor structure assessment in educational psychology research. *Frontiers in Education* (Vol. 4, p. 45).

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal, 15*(1), 136-153.

Kline, T. J. B. (2005). *Psychological testing: A practical approach to design and evaluation,* Thousand Oaks, CA: Sage.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 257-269). New York: Springer.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14*(1), 59-71.

Muraki, E. & Engelhard, G. Jr. (1985). Full-information item factor analysis: Applications of EAP scores. Applied Psychological Measurement, 9, 417-430.

Muthén , B.O, du Toit, S., & Spisic,, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes.* Unpublished manuscript.

Reckase, M. D. (2009). Multidimensional Item Response Theory. New York, NY: Springer.
Reise, S. P. (2012). The rediscovery of the bifactor measurement models. *Multivariate Behavioral Research, 47,* 667–696. doi: 0.1080/00273171.2012.715555
Reise, S. P., and Widaman, K. F. (1999). Assessing the fit of measurement models at the individual level: a comparison of item response theory and covariance structure approaches. *Psychological Methods, 4*, 3–21. doi: 10.1037/1082-989X.4.1.3
Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate behavioral research, 23*(1), 51-67.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplements, 17*.
Teresi JA, Jones RN. (2013). Bias in psychological assessment and other measures In Geisinger K, Bracken B, Carlson J, Hansen J-I, Kuncel N, Reise S, et al., editors. *APA handbook of testing and assessment in psychology, Vol 1: Test theory and testing and assessment in industrial and organizational psychology APA handbooks in psychology* (pp. 139–164).. Washington, DC: American Psychological Association
Thorpe, G. L. and Favia, A., (2012). *Data Analysis Using Item Response Theory Methodology: An Introduction to Selected Programs and Applications. Psychology Faculty Scholarship*. https://digitalcommons.library.umaine.edu/psy_facpub/20
Turk, D. C., Dworkin, R. H., Burke, L. B., Gershon, R., Rothman, M., Scott, J., et al. (2006). Developing patient-reported outcome measures for pain clinical trials: IMMPACT recommendations. *Pain, 125*, 208-215. doi:10.1016/j.pain.2006.09.028