**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 \( \theta \), 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 \( \theta \) 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 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 \( \alpha \) is the intercept from the right-hand side of the equation in the usual generalized linear model, then the difficulty parameter is \( b = \frac{\alpha}{a} \).

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.
One of the more common IRT models is the *normal ogive model*, which is essentially just the probit model (ogive 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 $\theta$ and the difficulty $b$, such that $\theta - b = 0$ and $\theta = b$, then the $z$-score is 0 and the probability 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.

$$z = a\theta - ab$$
$$y^* = \beta X - \alpha$$
$$= -\alpha + \beta X$$

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 $\theta$, and the intercept has a negative sign.

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.

![Normal ogive model](image1)

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.

![Normal ogive model](image2)

Raykov and Marcoulides (2011, p. 279)
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). 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 model can be extended to ordinal or multiple category responses using the other generalized linear regression models we have discussed.

Software

IRT models can be estimated with a variety of different software packages, such as stand alone packages like MULTILOG (Thissen, Chen, & Bock, 2002), ltm, TAM, and eRm packages in R, and the IRTFIT-RESAMPLE, DRAWICC, and IRGEN macros and procedures for SAS. 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).

References and Further Reading