ITEM RESPONSE THEORY

Randall E. Schumacker, Ph.D.

Introduction

Item response theory (IRT), based on latent trait theory, incorporates measurement assumptions about examinee item and test performance, and how performance relates to knowledge as measured by the items on a test. Within the general IRT framework, many models have been formulated. Popular names associated with these various scoring models are: dichotomous, binomial, Poisson, rating scale, facets, multinomial logit, or polytomous. These scoring models handle item responses that are discrete or continuous and dichotomous or polytomous scored.

One-, two-, and three-parameter IRT models

IRT models differ depending on whether the relationship between item performance and knowledge is considered a one-, two-, or three-parameter logistic function. Different IRT parameterization models adjust for different item properties leading to different ability estimation. 1-parameter (1-PL) IRT adjusts for item difficulty; 2-parameter (2-PL) IRT accounts for item difficulty and discrimination; and 3-parameter (3-PL) IRT takes into account the effect of item guessing, difficulty and discrimination. A popular one-parameter model, developed by George Rasch, is also commonly used where item difficulty provides an unbiased, efficient, sufficient, and consistent estimate of separate person and item calibrations.

The following figure depicts different item performance based on item difficulty and item discrimination. $P_i(\theta)$ is the probability of an examinee with ability θ answering item i correctly. The two parameters that characterize item i are a_i (the item discrimination), and b_i (the item difficulty). The normal ogive curve obtained by plotting $P_i(\theta)$ against θ is called an item characteristic curve (ICC). The b_i value is on the ability continuum where the $P_i(\theta) = 0.5$, and a_i is the slope of the curve at that point. The guessing parameter increases the starting point on the Y-axis for $P_i(\theta)$.

IRT Advantages

IRT measurement models, when compared to classical models, offer several distinct benefits. These include the following:

- Item statistics are independent of the sample from which they were estimated
- Examinee scores are independent of test difficulty
- Item analysis accommodates matching test items to examinee knowledge level
- Test analysis doesn't require strict parallel tests for assessing reliability
- Item statistics and examinee ability are both reported on the same scale.

IRT Disadvantages

IRT models have several technical and practical shortcomings. Assumptions underlying the use of IRT models are more stringent than those required of classical test theory. IRT models also tend to be more complex and the model-outputs more difficult to understand, particularly with non-technically oriented audiences. Additionally, IRT models require large samples to obtain accurate and stable parameter estimates, although Rasch measurement models are useful with small to moderate samples. Consequently, the choice of a model may depend upon the sample available, particularly in the field-testing phase of a certification exam.

Conclusion

In recent years, a definite trend has developed toward the use of IRT models by certification agencies. This is particularly true with the one-parameter Rasch model and the two-parameter IRT logistic model. This trend is primarily due to the objective measurement afforded by these models, coupled with linearity of scale, two features not found in classical theory. In addition, it is safe to assume that the movement toward item banking and computer adaptive testing will only serve to increase the popularity of IRT methodology. We are actively involved in assisting organizations in developing and maintaining IRT item banks, adaptive tests, and the use of various scoring models.