EPSY 8226: Item Response Theory

Scoring

Notation and definitions from Kolen, M.J., Tong, Y., & Brennan, R.L. (2011). Scoring and scaling educational tests. In A.A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking*. New York, NY: Springer.

**TRADITIONAL SCORING**

Item scores are combined in some way to create raw scores.

Scaling is the process of assigning numeric values to test taker performance as a result of educational or psychological tests.

A score scale is then produced to report scale scores for test takers.

Summed score X across *n* items

$$X=\sum\_{i=1}^{n}V\_{i}$$

Weighted summed score, Xw

$$X\_{w}=\sum\_{i=1}^{n}w\_{i}V\_{i}$$

**IRT SCORING**

Maximum Likelihood Scoring of $\hat{θ}\_{MLE}$

$$L=\prod\_{i=1}^{n}p(V\_{i}=v\_{i}|θ)$$

*Employing the Test Characteristic Function for Summed Scores*

The item response function employing summed scores is defined as:

$$τ\_{i}(θ)=\sum\_{k=1}^{m\_{i}}W\_{ik}∙P(V\_{i}=k|θ)$$

Where W*ik* is the score associated with item *i* and category *k*, this expresses the expected score on item *i* for a test taker with ability θ.

The test characteristics function is the sum of the item response functions across items:

$$τ(θ)=\sum\_{i=1}^{n}τ\_{i}(θ)$$

This provides the true score for a test taker with ability of θ. A weighted TCF can also be defined with *wi*. An ability estimate based on a summed score can be found by replacing $τ(θ)$ with the summed score and solving for θ.

The resulting score is $\hat{θ}\_{TCF}$.

*Note*: The estimates above do not depend on the population ability distribution. However, extreme scores do not exist. In addition, summed scores below the sum of the lower asymptote parameters do not exist. Bayesian estimators do exist in these cases, but they also depend on the population ability distribution.

**Bayesian Scoring with complicated pattern scoring**

With the specification of a population ability distribution (θ), we estimate $\hat{θ}\_{BME}$

$$L∙g(θ)=\prod\_{i=1}^{n}p(V\_{i}=v\_{i}|θ)∙g(θ)$$

For the expected a posteriori (EAP) estimator is the mean of the posterior distribution,

$\hat{θ}\_{EAP}$ = *E*(θ|V1=v1, V2=v2, …, Vn=vn)

**Bayesian Scoring using summed scores**

$\hat{θ}\_{sEAP}$ = *E*(θ|X)

This requires the use of a recursive algorithm that estimates the probability of obtaining a given summed score given ability, calculated from item parameter estimates.

***Linear Transformations***

It is possible to use linear transformations by directly linearly transforming IRT ability estimates. Nonlinear transformations are also possible, but the transformation is complicated for Bayesian estimates.

To transform raw scores (*y*) to a desired scale score (*sc*) distribution:

$$sc\left(y\right)=\frac{σ(sc)}{σ(Y)}y+\left[μ\left(sc\right)-\frac{σ\left(sc\right)}{σ\left(Y\right)}μ(Y)\right] or =σ(sc)z\_{Y}+μ\left(sc\right)$$

Linearly transformed score scales based on the standard error of measurement can be created to support more appropriate score interpretation:

$$sc\left(y\right)=\frac{sem\_{sc}}{sem\_{y}}y+\left[sc(y\_{1})-\frac{sem\_{sc}}{sem\_{y}}y\_{1}\right]$$

In this case, let *y*1 be a raw score and *sc*(*y*1) be its scale score equivalent, where *semsc* is the desired average scale score standard error of measurement, *semy* is the average raw score standard error of measurement.

**Incorporating Content Information**

*Item Mapping*

To enhance the meaning of the scale scores, items are found that represent various scale score points, and these representative items are reported to test users. …One choice made in item mapping is the response probability (RP) level, which is the probability of a correct response that is associated with mastery for all items on a test, expressed as a percentage. The mastery level for a specific item is defined as the scale score for which the probability times 100 of correctly answering the item equals the RP level. (p. 414)

*Scale Anchoring*

The goal of scale anchoring is to provide general statements of what students, who score at each of a selected set of scale score points, know and are able to do. Subject matter experts review the items that map near each point and attempt to develop general statements that represent the skills of examinees scoring at these points. In scale anchoring, it is assumed that examinees know and are able to do all of the skills in the statements at or below a given score level. (p. 415)

*Standard Setting*

Standard setting begins with a statement of what competent examinees know and are able to do. Standard setting methods are an attempt to find the score point that divides those examinees who know and are able to do what is stated from the other examinees. (p. 417)