# 3. OVERVIEW OF STATISTICAL SUMMARIES

This section provides general information about statistical and psychometric summaries used for the 2003 *MSA-Reading* program. Actual statistical results described in this section appear in section 4 and Appendices.

## 3.1 Classical Descriptive Statistics

Table 4.1 contains the classical descriptive statistics of each form for each grade and includes:

- Form number
- Number of items
- Numbers of students[1]
- Means and standard deviations of raw scores
- *Stratified Cronbach Alpha*
- *Standard error of measurement* (*SEM*)

**Stratified Cronbach Alpha**

The 2003 *MSA-Reading* tests included both *SR* and *BCR* items. Consequently, it was asked to use an adequate reliability coefficient that addressed the important factor, different item type. The following formula depicts the reliability coefficient, *Stratified Cronbach Alpha*:

$$Stratified\ a = 1 - \frac{\sum s_i^2 (1 - r_{ii'})}{s_t^2}$$

where

$s_i^2$ = variance of score on cluster *i*,

$s_t^2$ = variance of total score, and

$r_{ii'}$ = reliability coefficient of score on cluster *i*.

These tests were initially considered to be *classically congeneric* (i.e., besides having unequal means and unequal variances in error and observed scores, the test forms also have heterogeneity of true-score variances) where the tasks within the examinations were stratified based on the type of item (i.e., multiple-choice, short answers, extended responses, and extended writing) and by the scoring rubric attached to these items.

Upon examining the variance/covariance matrices, however, it became apparent that in some cases the part covariance of a part was not heterogeneous with respect to other part variances (e.g., the covariance between multiple-choice items and extended responses and between

[1] Note the numbers of students reported in these tables may be lower than the totals reported in the statewide summaries. These analyses were based on the sample of data used to equate the forms of the 2003 *MSA-Reading*.

multiple-choice items and extended writing for grade 3 writing). It was, therefore, determined that although the test forms were *congenerically parallel*, they were not *classically congeneric* (Qualls, 1995). For the 2003 *MSA-Reading*, therefore, the test forms were divided into two strata made up of *SR* and *BCR* items, and the *Stratified Cronbach Alpha* was used as the reliability coefficient.

**Standard Error of Measurement (Based on Classical Test Theory)**

The *standard error of measurement* (*SEM*) is the standard deviation of errors of measurement that are associated with test scores from a particular group of examinees. In here, a measurement error is the difference between an examinee's actual or obtained score and the theoretical true score counterparts. Consequently, the *SEM* is commonly used in interpreting and reporting individual test scores and score differences on tests (Harvill, 1991).

Classical test theory is based on the following assumptions (Andrich & Luo, 2003):

- Each person *v* has a true score on the construct, usually denoted by the variable $T_v$
- The best overall indicator of the person's true score is the sum of the scores on the items and is usually denoted by the variable $X_v$
- This observed score will have an error for each person which is usually denoted by $E_v$
- These errors are not correlated with the true score
- Across a population of people, the errors sum to 0 and they are normally distributed.

From these assumptions, the following equations can be derived:

$$X_v = T_v + E_v .$$

Therefore,

$$S_x^2 = S_t^2 + S_e^2$$

where

$S_x^2$ = the variance of the observed score in a population of persons,

$S_t^2$ = the variance of their true score variance, and

$S_e^2$ = the error variance.

The reliability of the test can be calculated by the following formula:

$$r_{xx} = \frac{s_t^2}{s_x^2} = \frac{s_x^2 - s_e^2}{s_x^2} .$$

Thus, the *SEM* is calculated by the following formula:

$$S_e = S_x\sqrt{1 - r_{xx}} .$$

For example, consider a student with a score of 90 from a sample of students with a mean score of 60 and variance of 225 on a test with reliability of 0.80. According to the formulas provided above, the obtained score is 90, and its *SEM* is 6.71. Thus, an approximate 68% score band for estimating this students' true score is from 83.29 (90 - 6.71) to 96.71 (90 + 6.71).

Note that this equation is only useful to estimate true score when the test reliability is reasonably high and the obtained score for the examinee is not an extreme deviate from the mean of the appropriate reference group. When we use this equation, consequently, we should be careful with statements so that they do not imply greater precision than is actually involved (Harvill, 1991).

**Conditional Standard Error of Measurement (Based on Item Response Theory)**

Unlike the *SEM* based on the classical test theory, the *SEM* based on the *item response theory* (*IRT*) is not the same for all persons. For example, if a person gets few or a large number of items correct, the standard error is greater than if the person gets moderate number of items correct. This implies that the standard error of measurement depends on the total score (Andrich & Luo, 2003).

Under the Rasch model, the *SEM* for each person is as follows:

$$s_{\hat{b}} = \frac{1}{\sqrt{\sum_{i=1}^{L} p_{vi}(1-p_{vi})}}$$

where

v = subscript for a person,

i = subscript for an item,

L = length of the test,

$\hat{b}$ = ability estimate, and

$p_{vi}$ = the probability that a person answers an item correctly and defined as follows:

$p_{vi} = \frac{e^{b_v - d_i}}{1 + e^{b_v - d_i}}$ where $b_v$ is person's ability and $d_i$ is item's difficulty.

A confidence band can be found for use in interpreting the ability estimate. For example, an approximate 68% confidence interval for $\hat{b}$ is given by

$\hat{b} \pm SEM$

Note that the standard error for item difficulty is smallest when the probability of passing is close to the probability of failing. That is, when an item is near the threshold level for many persons in the sample, the standard error is small (Embretson & Reise, 2000).

## 3.2 Scale Score Descriptive Statistics

Table 4.2 provides information about scale score descriptive statistics of each form for each grade and includes:

- Form number
- Number of items
- Numbers of students
- Means and standard deviations of scale scores
- Conditional *standard errors of measurement* (*SEM*) for the proficient and advanced cut scores

In addition, Appendix A provides frequency distributions and histograms of the scale scores of the 2003 *MSA-Reading*.

## 3.3 Classical and IRT Item Parameters

Appendix B provides both classical and *IRT*-based item parameters and includes:

- Item type (*SR* or *BCR*)
- *p*-value
- Point-biserial correlation: in order for *p*-values of the *BCR* items to be comparable with *p*-values of the *SR* items they were calculated as modified proportions of the maximum obtainable domain scores.
- Rasch difficulty estimate
- Standard error of the Rasch difficulty
- Mean-square infit
- Mean-square outfit

Item sequence numbers represents merely those items that were chosen to be in the final "score form."

**Fit Statistics for Rasch Model**

Fit statistics are used for evaluating the goodness-of-fit of a model to the data. Fit statistics are calculated by comparing the observed and expected trace lines obtained for an item after parameter estimates are obtained using a particular model. *WINSTEPS* provides two kinds of fit statistics called *mean-squares* that show the size of the randomness or amount of distortion of the measurement system.

*Outfit* mean-squares are influenced by outliers and are usually easy to diagnose and remedy. *Infit* mean-squares, on the other hand, are influenced by response patterns and are harder to diagnose and remedy. Table 3.1 provides a guideline for evaluating mean-square fit statistics (Linacre & Wright, 2000).

In general, mean-squares near 1.0 indicate little distortion of the measurement system, while values less than 1.0 indicate observations are too predictable (redundancy, model overfit). Values greater than 1.0 indicate unpredictability (unmodeled noise, model underfit).

**Table 3.1 Criteria to Evaluate Mean-Square Fit Statistics**

| Mean-Square | Interpretation |
|---|---|
| > 2.0 | Distorts or degrades the measurement system |
| 1.5 – 2.0 | Unproductive for construction of measurement, but not degrading |
| 0.5 – 1.5 | Productive for measurement |
| < 0.5 | Unproductive for measurement, but not degrading. May produce misleadingly good reliabilities and separations |

## 3.4 Inter-Rater Reliability

Tables 4.30 through 4.32 contain information about the scoring agreement between rater 1 and rater 2. When the two readers assigned the same score to a student's answer, the scores were in perfect agreement. Scores differed by one score point were adjacent, and scores differed by two or more score points were in discrepancy. For further information about inter-rater agreement, please see section 1.7. For the 2003 *MSA-Reading*, the adjacent agreement rates were above 95%, and perfect agreement rates were above 70% except several items for each grade.

## 3.5 Correlations among Reading Processes

The 2003 *MSA-Reading* consisted of three reading processes (strands): *General Reading*, *Literary Reading*, and *Informational Reading*. Tables 4.3 through 4.5 contain correlation coefficients among these reading processes. Generally, they show moderately strong intercorrelations among them.

## 3.6 Decision Accuracy and Consistency at the Cut Scores

Tables 4.6 through 4.8 contain the results of analyses performed to estimate the accuracy and consistency of the decisions for passing (proficient) on the 2003 *MSA-Reading*. The analyses make use of the methods outlined and implemented in Livingston and Lewis (1995), Haertel (1996), and Young and Yoon (1998).

The *accuracy* of a decision is the extent to which it would agree with the decisions that would be made if each student could somehow be tested with all possible parallel forms of the assessments. The *consistency* of a decision is the extent to which it would agree with the decisions that would be made if the students had taken a different form of the examination, equal in difficulty and covering the same content as the form they actually took.

Students can be misclassified in one of two ways. Students who were below the proficiency cut score, but were classified (on the basis of the assessment) as being above a cut score, are considered to be *false positives*. Students who were above the proficiency cut score, but were classified as being below a cut score, are considered to be *false negatives*.

For the 2003 *MSA-Reading*, Tables 4.6 through 4.8 include:

- Performance level
- Accuracy classifications
- False positives
- False negatives
- Consistency classifications

The tables illustrate the general rule that decision consistency is less than decision accuracy.

## 3.7 Differential Item Functioning

This section provides information about *differential item functioning* (*DIF*) analyses used for the 2003 *MSA-Reading*. For the 2003 *MSA-Reading DIF* analyses, the *reference* group was either male or Caucasian students, and the *focal* group was either female or African-American students.

Since the 2003 *MSA-Reading* was a mixed-format examination, comprising of both *BCR* and *SR* items, the *DIF* procedure used consists of Mantel's (1963) extension of the Mantel-Haenszel procedure (the Mantel Chi-square) for the *BCR* items and the Mantel-Haenszel procedure (Mantel & Haenszel, 1959) for the *SR* items.

**Brief Constructed Response (BCR) Items**

To help interpret the Mantel Chi-square (Mantel $?^2$), the Educational Testing Service (ETS) *DIF* procedure uses the Mantel statistic in conjunction with the *standardized mean difference* (*SMD*).

**Mantel Statistic**

The Mantel $?^2$ is simply a conditional mean comparison of the ordered response categories for reference and focal groups combined over values of the matching variable score. By "ordered" we mean that a response of 1 on an item is better than 0, 2 is better than 1, and so on. "Conditional," on the other hand, refers to the comparison of members from the two groups who received the same score on the matching variable, i.e., the total test score in our analysis.

Table 3.2 shows a $2 \times T \times K$ contingency table, where $T$ is the number of response categories and $K$ is the number of levels of the matching variable. The values, $y_1, y_2, \ldots, y_T$ are the $T$ scores that can be gained on the item. The values, $n_{Ftk}$ and $n_{Rtk}$, represent the numbers of focal and reference groups who are at the $k^{th}$ level of the matching variable and gain an item score of $y_t$. The "+" indicates total number over a particular index (Zwick, Donoghue, & Grima, 1993).

**Table 3.2 $2 \times T$ Contingency Table at the $k^{th}$ level ***

| Group | Item Score | | | | Total |
|---|---|---|---|---|---|
| | $y_1$ | $y_2$ | | $y_T$ | |
| Reference | $n_{R1k}$ | $n_{R2k}$ | … | $n_{RTk}$ | $n_{R+k}$ |
| Focal | $n_{F1k}$ | $n_{F2k}$ | … | $n_{FTk}$ | $n_{F+k}$ |
| Total | $n_{+1k}$ | $n_{+2k}$ | … | $n_{+Tk}$ | $n_{++k}$ |

*Zwick, et al. (1993)

The Mantel statistics is defined as the following formula:

$$\text{Mantel } ?^2 = \frac{(\sum_k F_k - \sum_k E(F_k))^2}{\sum_k Var(F_k)}$$

where

$F_k$ = the sum of scores for the focal group at the $k^{th}$ level of the matching variable and is defined as follows:

$$F_k = \sum_t y_t n_{Ftk},$$

The expectation of $F_k$ under the null hypothesis is

$$E(F_k) = \frac{n_{F+k}}{n_{++k}} \sum_t y_t n_{+tk}.$$

And, the variance of $F_k$ under the null hypothesis is as follows:

$$Var(F_k) = \frac{n_{R+k} n_{F+k}}{n_{++k}^2 (n_{++k} - 1)} \left[ (n_{++k} \sum_t y_t^2 n_{+tk}) - (\sum_t y_t n_{+tk})^2 \right].$$

Under $H_0$, the Mantel statistic has a chi-square distribution with one degree of freedom. In *DIF* applications, rejecting $H_0$ suggests that the students of the reference and focal groups who are similar in overall test performance tend to differ in their mean performance. In the case of dichotomous items, on the other hand, the statistics is identical to the Mantel-Haenszel (1959) statistic without the continuity correction (Zwick, Donoghue, & Grima, 1993).

**Standardized Mean Difference (SMD)**

A summary statistic to accompany the Mantel approach is the *standardized mean difference* (*SMD*) between the reference and focal groups proposed by Dorans and Schmitt (1991). This statistic compares the means of the reference and focal groups, adjusting for differences in the distribution of the reference and focal group members across the values of the matching variable.

$$SMD = \sum_k p_{Fk} m_{Fk} - \sum_k p_{Fk} m_{Rk}$$

where

$p_{Fk} = \frac{n_{F+k}}{n_{F++}}$, the proportion of the focal group members who are at the $k^{th}$ level of the matching variable,

$m_{Fk} = \frac{1}{n_{F+k}(\sum_t y_t n_{Ftk})}$, the mean item score of the focal group members at the $k^{th}$ level, and

$m_{Rk}$ = the analogous value for the reference group.

As can be seen from the equation above, the *SMD* is the difference between the unweighted item mean of the focal group and the weighted item mean of the reference group. The weights for the reference group are applied to make the weighted number of the reference group students the same as in the focal group within the same ability. A negative *SMD* value implies that the focal group has a lower mean item score than the reference group, conditional on the matching variable.

**DIF classification for BCR items**

The *SMD* is divided by the total group item standard deviation to obtain an effect-size value for the *SMD*. This effect-size *SMD* is then examined in conjunction with the Mantel $?^2$ to obtain *DIF* classifications that are depicted in Table 3.3 below.

**Table 3.3 *DIF* Classification for BCR Items**

| Category | Description | Criterion* |
|---|---|---|
| AA | No *DIF* | Non-significant Mantel $?^2$ or<br>Significant Mantel $?^2$ and \|SMD/SD\| = .17 |
| BB | Weak *DIF* | Significant Mantel $?^2$ and .17 < \|SMD/SD\| = .25 |
| CC | Strong *DIF* | Significant Mantel $?^2$ and .25 < \|SMD/SD\| |

* SD is the total group standard deviation of the item score in its original metric

**Selected Response (SR) Items**

For the *SR* items, the Mantel-Haenszel Chi-square (M-H $?^2$) in conjunction with the M-H odds ratio that is transferred to what ETS calls, the delta scale (D).

**The Odds Ratio**

The odds of a correct response (proportion passing divided by proportion failing) are *P/Q* or *P/(1-P)*. The odds ratio, on the other hand, is simply the odds of a correct response of the reference group divided by the odds of a correct response of the focal group.

For a given item, the odds ratio is defined as follows:

$$a_{M-H} = \frac{P_r / Q_r}{P_f / Qf}.$$

And, the corresponding null hypothesis is that the odds of getting the item correct are equal for the two groups. Thus, the odds ratio is equal to 1:

$$H_0: a_{M-H} = \frac{P_r / Q_r}{P_f / Qf} = 1.$$

## The Delta Scale

In order to make the odds ratio symmetrical around zero with its range being in the interval $-\infty$ to $+\infty$, the odds ratio is transformed into a log odds ratio as per the following:

$$b_{M-H} = \ln(a_{\text{M-H}}).$$

The simple natural logarithm transformation of this odds ratio is symmetrical about zero in which zero has the interpretation of equal odds. This *DIF* measure is a signed index where a positive value signifies *DIF* in favor of the reference group while a negative value indicates *DIF* in favor of the focal group. $b_{M-H}$ also has the advantage of being transformed linearly to other interval scale metrics (Camilli & Shepard, 1994). This fact is utilized by ETS in creating their delta scale (D), which is defined as follows:

$$\text{D} = -2.35 \cdot b_{M-H}.$$

## DIF classification for SR items

The ETS examines the M-H $?^2$ in conjunction with the delta scale (D) to obtain *DIF* classifications depicted in Table 3.4 below.

**Table 3.4 *DIF* Classification for SR Items**

| Category | Description | Criterion |
|---|---|---|
| A | No *DIF* | Non-significant M-H $?^2$ or \|D\| < 1.0 |
| B | Weak *DIF* | Significant M-H $?^2$ and \|D\| < 1.5 or<br>Non-significant M-H $?^2$ and \|D\| =1.0 |
| C | Strong *DIF* | Significant M-H $?^2$ and \|D\| = 1.5 |

## 3.8 Equating and Scaling

Tables 4.9 through 4.29 contain information about raw score to scale score conversion tables for the 2003 *MSA-Reading*. Conditional standard errors for the scale scores are also included.

### The Rasch and Partial Credit IRT Models

The most basic expression of the Rasch model is in the *item characteristic curve* (ICC). It shows the probability of a correct response to an item as a function of the ability level. The probability of a correct response is bounded by 1 (certainty of a correct response) and 0 (certainty of an incorrect response). The ability is, in theory, unbounded. In practice, the ability scale ranges from - 4 to + 4 logits for heterogeneous ability groups.

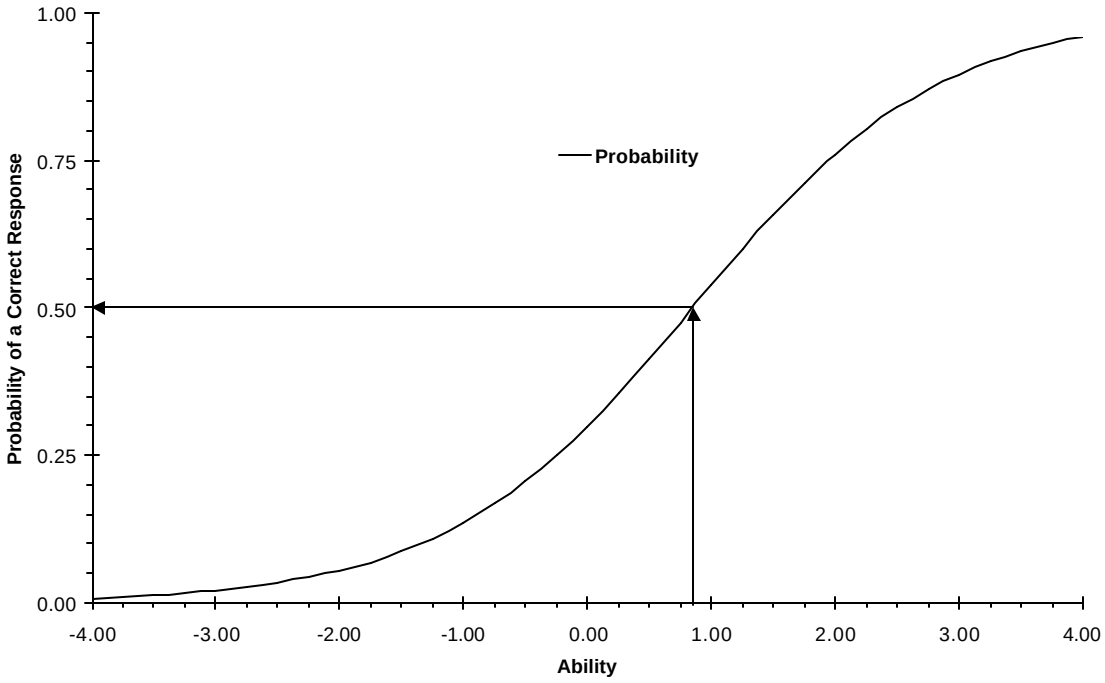
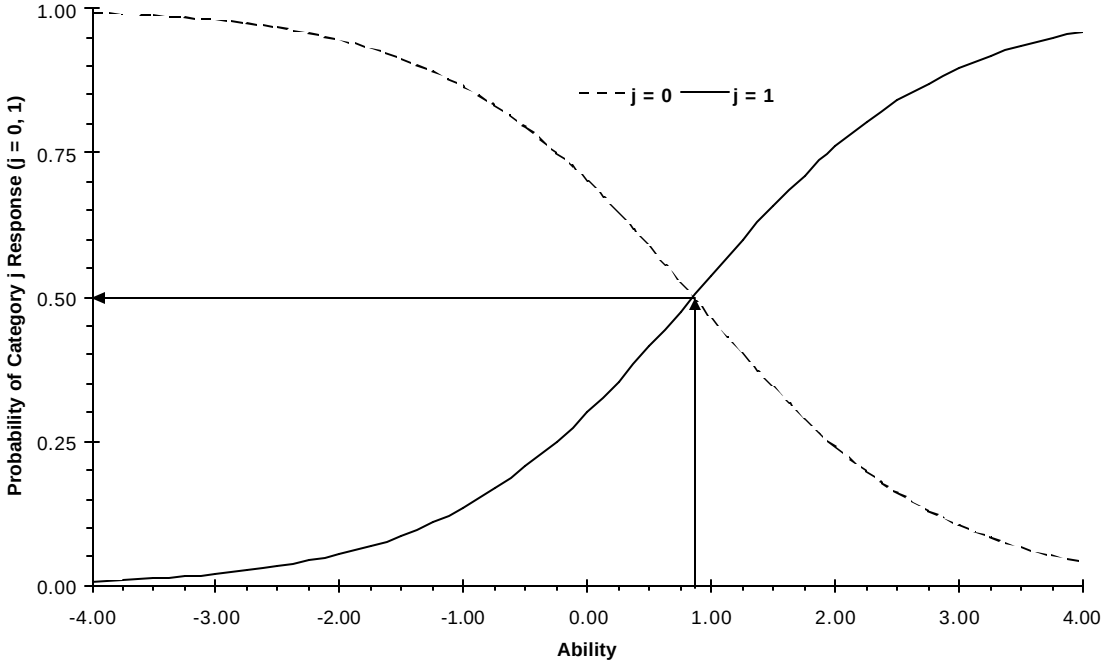



**Figure 3.1 Item Characteristic Curve**

As an example, consider Figure 3.1 which depicts a item that falls at approximately 0.85 on the ability (horizontal) scale. When a person answers an item at the same level as their ability, then that person has a probability of roughly 50% of answering the item correctly. Another way of expressing this is that if we have a group of 100 people, all of whom have an ability of 0.85, we would expect about 50% of them to answer the item correctly. A person whose ability was above 0.85 would a higher probability of getting the item right, while a person whose ability is below 0.85 would have a lower probability of getting the item right. This makes intuitive sense and is the basic formulation of Rasch measurement for test items having only 2 possible categories (i.e., wrong or right).




**Figugure 3.2 Category Response Curves for a One-Step Item**

Figure 3.2 extends this formulation to show the probabilities of obtaining a wrong answer or a right answer. The curve on the left ($j = 0$) shows the probability of getting a score of "0" while the curve on the right ($j = 1$) shows the probability of getting a score of "1". The point at which the two curves cross indicates the transition point on the ability scale where the most likely response changes from a "0" to a "1". Here, the probability of answering the item correctly is 50%.

The key step in the formulation, and the point at which the Rasch dichotomous model merges with the PCM, requires us to assume an additional response category. Suppose that, rather than scoring items as completely wrong or completely right, we add a category representing answers that, though not totally correct, are still clearly not totally incorrect. These relationships are shown in Figure 3.3.

The left-most curve ($j = 0$) in Figure 3.3 represents the probability for all examinees getting a score of "0" (completely incorrect) on the item, given their ability. Those of very low ability (e.g., below - 2) are very likely to be in this category and, in fact, are more likely to be in this category than the other two. Those receiving a "1" (partial credit) tend to fall in the middle range of abilities (the middle curve, $j = 1$). The final, right-most curve ($j = 2$) represents the probability for those receiving scores of "2" (completely correct). Very high-ability people are clearly more likely to be in this category than in any other, but there are still some of average and low ability that can get full credit for the item.

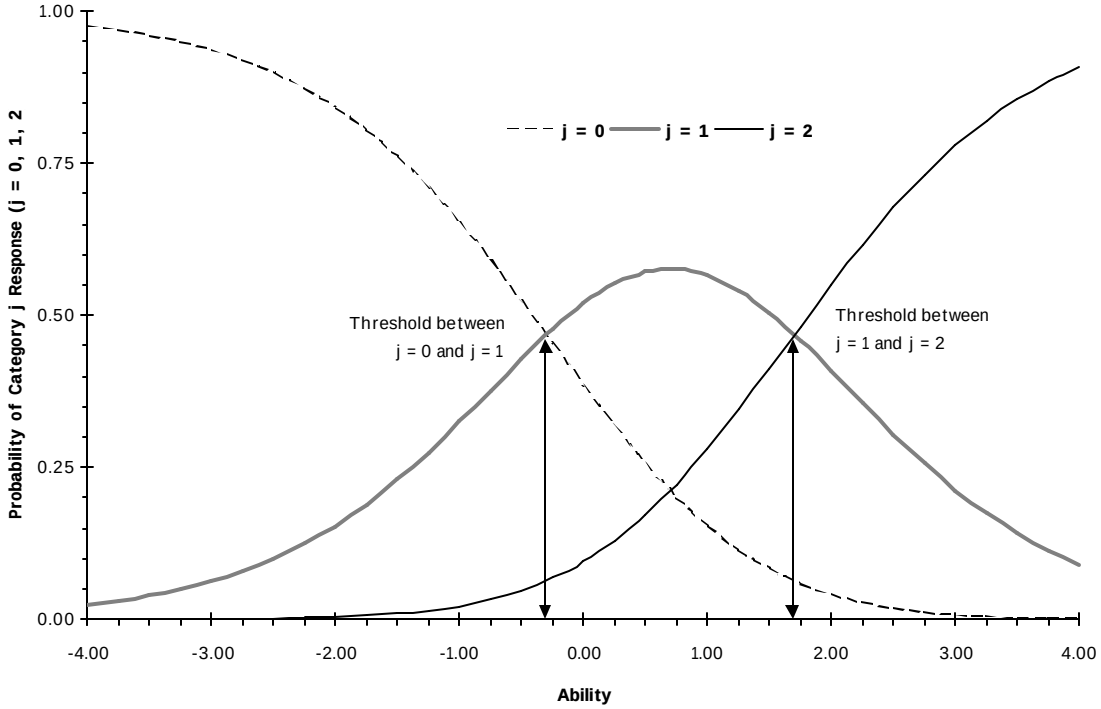



**Figure 3.3 Category Response Curves for a Two-Step Item**

Although the actual computations are quite complex, the points at which lines cross each other have a similar interpretation as for the dichotomous case. Consider the point at which the $j = 0$ line crosses the $j = 1$ line, indicated by the left arrow. For abilities to the left of (or less than) this point, the probability is greatest for a "0" response. To the right of (or above) this point, and up to the point at which the $j = 1$ and $j = 2$ lines cross (marked by the right arrow), the most likely response is a "1". For abilities to the right of this point, the most likely response is a "2".

Note that the probability of scoring a "1" response ($j = 1$) declines in both directions as ability decreases to the low extreme or increases to the high extreme. These points then may be thought of as the difficulties of crossing the *thresholds* between categories.

An important implication of the formulation can be summarized as: If the commonly used Rasch model applied to dichotomously (right/wrong) scored items can be thought of as simply a special case of the PCM, then the act of scaling multiple-choice items together with polytomous items, whether they have three or more response categories, is a straightforward process of applying the measurement model. The quality of the scaling then can be assessed in terms of known procedures.

One important property of the PCM is its ability to separate the estimation of item/task parameters from the person parameters. With the PCM, as with the Rasch model, the total score given by the sum of the categories in which a person responds is a sufficient statistic for estimating person ability (i.e., no additional information need be estimated). The total number of

responses across examinees in a particular category is a sufficient statistic for estimating the step difficulty for that category. Thus with PCM, the same total score will yield the same ability estimate for different examinees.

The PCM is a direct extension of the dichotomous one-parameter IRT model developed by Rasch (Rasch, 1980). For an item/task involving $m_i$ score categories, one general expression for the probability of scoring $x$ on item/task $i$ is given by

$$P_{xi} = \exp \sum_{j=0}^{x} (q - D_{ij}) / \sum_{k=0}^{m_i} \left[ \exp \sum_{j=0}^{k} (q - D_{ij}) \right]$$

where

$x = 0, 1, ..., m_i$, and by definition, $\sum_{j=0}^{0} (q - D_{ij}) = 0$.

The above equation gives the probability of scoring $x$ on the $i$-th test item as a function of ability ($q$) and the difficulty of the $m_i$ steps of the task (Masters, 1982).

According to this model, the probability of an examinee scoring in a particular category (step) is the sum of the logit (log-odds) differences between $q$ and $D_{ij}$ of all the completed steps, divided by the sum of the differences of all the steps of a task. Thissen and Steinberg (1986) refers to this model as a divide-by-total model. The parameters estimated by this model are (1) an ability estimate for each person (or ability estimate at each raw score level) and (2) $m_i$ threshold (difficulty) estimates for each task with $m_i + 1$ score categories.