Wisconsin Forward Exam: Standard Performance Index

August 10, 2016

Overview of Standard Performance Index Score Concept and Use

In addition to total test raw scores and scale scores, teachers and educational decision-makers frequently seek information to inform instructional strategies. This information can be derived from scores on subsets of test items that estimate how much a student knows in a clearly defined skill domain. These skill domains are called content standards (or standards or objectives). Scores on subsets of test items at the content standard level are called standard performance index (SPI) scores. The purpose of reporting SPI scores on the Wisconsin Forward Exam is to show the relationship between the overall achievement being measured (represented by the test score) and the skills within each of the content standards associated with the content area. Teachers may use the SPI scores for individual students as indicators of strengths and needs, but the SPI scores are best corroborated by other evidence, such as homework, class participation, diagnostic test scores, or observation. District and school administrators may compare their results by content standard and grade level with the state mean percentage to better understand their strengths and needs within a particular content area and grade level.

SPI scores range from 1 to 99 and can be interpreted as an estimate of the number of items a student would be expected to answer correctly if there had been 100 similar items for a given reporting category. For example, an SPI of 77 for a given reporting category means that if the student was given 100 similar items, the student would be expected to answer 77 of them correctly. These are criterion referenced scores, in that they estimate how much a student knows in a clearly defined skill domain (i.e., the criterion). The SPI scores are computed for content standards measured by at least 4 items. The SPI scores are not computed for content standards measured by fewer than 4 items.

Based on their SPI scores, students are classified in one of the four content category performance levels: below basic, basic, proficient, and advanced. The SPI cut scores separating these performance levels are derived as expected percentages of possible score points for a given standard (content category) for students whose total test score is at the corresponding total test cut score (basic, proficient, or advanced).

Because a student's overall achievement is factored in computation of the SPI scores, these scores provide more reliable estimates of student achievement on each content standard than simple percent correct. However, *the SPI information should only be used for low-stakes purposes because the SPI cannot be considered stable for any content standard with a small number of items.* The average difficulty of items may vary across content standards and grades. Content standards vary in their complexity, level of abstraction, and cognitive demand. Some standards may be intrinsically more difficult than others, and the difficulty of individual items is determined, in part, by the difficulty of the content domain being measured. The current test blueprints do not specify the average difficulty level of items for each content standard within grades or across grades. If the difficulty of the items varies across years, grades, and content standards, the mean SPI scores will be affected by differences in item difficulty as well as differences in student ability. *Thus, differences in SPI scores across years, grades, or content standards should not be seen as reliable indicators of differences in student ability since these differences may be explained in whole or in part by differences in the difficulty of the items themselves.* However, comparisons across years, grades, or content standards are appropriate for assessing the relative difficulty of the items, and comparisons of individual student scores or of group mean scores on a single SPI can provide useful information about the *relative* strengths and needs of individual students or groups on these standards.

Technical Details of SPI Score Computation

Technical details of the SPI estimation procedure described in this section are based on description of the SPI computation methodology included in the *TerraNova* 2nd Edition Technical Report (CTB/McGraw-Hill, 2000).

The Standard Performance Index (SPI) is an estimate of the true score (estimated proportion of total, or maximum, points possible) for a content standard based on the performance of a given student. Because most standards are measured by a relatively small number of items, a Bayesian procedure that takes into account the overall test performance is used to improve the reliability of the standard scores. Given a student's scale score on the test, item response theory (IRT) is used, via the 3-paremeter logistic (3PL) model for MC items and the 2-paremeter-partial credit (2PPC) model for CR items, to compute the estimated proportion of the maximum points obtained for that standard.

The estimated proportion of the maximum points obtained for the standard provides the initial (Bayesian prior) estimate of the student's mastery score. If this initial estimate is consistent with the student's observed proportion, as indicated by a chi-square test, the two scores are combined as a weighted average to obtain the SPI score (the estimated true score). The appropriate weight for the Bayesian prior estimate is computed as a function of the standard error (SE) of the scale score on which it is based: the smaller the SE, the larger the weight. If the prior estimate and the observed proportion differ significantly, the observed proportion of the maximum score is used without the prior estimate to compute the student's score on that objective.

Standard Performance Index Computation

The standard performance index (SPI) is an estimated true score (estimated proportion of total or maximum points obtained) based on the performance of a given examinee for the items in a given learning strand. Assume a *k*-item test is composed of *j* strands with a maximum possible raw score of *n*. Also assume that each item contributes to, at most, one strand, and the k_j items in strand *j* contribute a maximum of n_j points. Define X_j as the observed raw score on strand ***j***. The true score is

$$T_j \equiv E(X_j / n_j).$$

It is assumed that there is information available about the examinee in addition to the strand score, and this information provides a prior distribution for T_j. This prior distribution of T_j for a given examinee is assumed to be $\beta(r_j, s_j)$:

$$g(T_j) = \frac{(r_j + s_j - 1)!\, T_j^{\,r_j - 1}(1 - T_j)^{s_j - 1}}{(r_j - 1)!(s_j - 1)!} \qquad (1)$$

for $0 \le T_j \le 1;\ r_j, s_j > 0$. Estimates of r_j and s_j are derived from IRT (Lord, 1980).

It is assumed that X_j follows a binomial distribution, given T_j:

$$p(X_j = x_j \mid T_j) = Binomial\ (n_j, T_j = \sum_{i=1}^{k_j} T_i / n_j),$$

where

T_i is the expected value of the score for item *i* in strand *j* for a given θ.

Given these assumptions, the posterior distribution of T_j, given x_j, is

$$g(T_j \mid X_j = x_j) = \beta(p_j, q_j), \qquad (2)$$

with

$$p_j = r_j + x_j \qquad (3)$$

and

$$q_j = s_j + n_j - x_j. \qquad (4)$$

The SPI is defined to be the mean of this posterior distribution:

$$\tilde{T}_j = \frac{p_j}{p_j + q_j}.$$

Following Novick and Jackson (1974, p. 119), a mastery band is created to be the *C*% central credibility interval for T_j. It is obtained by identifying the values that place $\frac{1}{2}(100-C)\%$ of the $\beta(p_j, q_j)$ density in each tail of the distribution.

Estimation of the Prior Distribution of T_j

The *k* items in each test are scaled together using a generalized IRT model (3PL/2PPC) that fits a three-parameter logistic model (3PL) to the MC items and a generalized partial-credit model (2PPC) to the CR items (Yen, 1993).

The 3PL model is

$$P_i(\theta) = P(X_i = 1 \mid \theta) = c_i + \frac{1-c_i}{1+\exp\left[-1.7A_i\left(\theta - B_i\right)\right]} \quad , \tag{5}$$

where

A_i is the discrimination, B_i is the location, and c_i is the guessing parameter for item *i*.

A generalization of Master's (1982) partial credit (2PPC) model was used for the CR items. The 2PPC model, the same as Muraki's (1992) "generalized partial credit model," has been shown to fit response data obtained from a wide variety of mixed-item type achievement tests (Fitzpatrick, Link, Yen, Burket, Ito, & Sykes, 1996). For a CR item with 1_i score levels, integer scores were assigned that ranged from 0 to $1_i - 1$:

$$P_{im}(\theta) = P(X_i = m-1|\theta) = \frac{\exp(z_{im})}{\sum_{g=1}^{1_i}\exp(z_{ig})}, \qquad m = 1, \ldots 1_i \tag{6}$$

where

$$z_{ig} = \alpha_i\left(m-1\right)\theta - \sum_{h=0}^{m-1}\gamma_{ih}, \tag{7}$$

and

$\gamma_{i0} = 0$.

Alpha (α_i) is the item discrimination, and gamma (γ_{ih}) is related to the difficulty of the item levels; the trace lines for adjacent score levels intersect at γ_{ih}/α_i .

Item parameters estimated from the national standardization sample are used to obtain SPI values. $T_{ij}(\theta)$ is the expected score for item *i* in strand *j*, and θ is the common trait value to which the items are scaled:

$$T_{ij}(\theta) = \sum_{m=1}^{1_i} (m-1) P_{ijm}(\theta),$$

where

1_i is the number of score levels in item *i*, including 0.

T_j, the expected proportion of maximum score for strand *j*, is

$$T_j = \frac{1}{n_j}\left[\sum_{i=1}^{k_j} T_{ij}(\theta)\right]. \quad (8)$$

The expected score for item *i* and estimated proportion-correct of maximum score for strand *j* are obtained by substituting the estimate of the trait $(\hat{\theta})$ for the actual trait value.

The theoretical random variation in item response vectors and resulting $(\hat{\theta})$ values for a given examinee produces the distribution $g(\hat{T}_j \mid \hat{\theta})$ with mean $\mu(\hat{T}_j \mid \theta)$ and variance $\sigma^2(\hat{T}_j \mid \theta)$. This distribution is used to estimate a prior distribution of T_j. Given that T_j is assumed to be distributed as a beta distribution (equation 1), the mean $[\mu(\hat{T}_j \mid \theta)]$ and variance $[\sigma^2(\hat{T}_j \mid \theta)]$ of this distribution can be expressed in terms of its parameters, r_j and s_j.

Expressing the mean and variance of the prior distribution in terms of the parameters of the beta distribution (Novick & Jackson, 1974, p. 113) produces

$$\mu(\hat{T}_j \mid \theta) = \frac{r_j}{r_j + s_j} \quad (9)$$

and

$$\sigma^2(\hat{T}_j \mid \theta) = \frac{r_j s_j}{(r_j + s_j)^2 (r_j + s_j + 1)}. \quad (10)$$

Solving these equations for r_j and s_j produces

$$r_j = \mu(\hat{T}_j \mid \theta) n_j^* \quad (11)$$

and

$$s_j = [1 - \mu(\hat{T}_j \mid \theta)] n_j^*, \quad (12)$$

where

$$n_j^* = \frac{\mu(\hat{T}_j \mid \theta)\left[1-\mu(\hat{T}_j \mid \theta)\right]}{\sigma^2(\hat{T}_j \mid \theta)} - 1. \tag{13}$$

Using IRT, $\sigma^2(\hat{T}_j \mid \theta)$ can be expressed in terms of item parameters (Lord, 1983):

$$\mu(\hat{T}_j \mid \theta) \approx \frac{1}{n_j} \sum_{i=1}^{k_j} \hat{T}_{ij}(\theta). \tag{14}$$

Because T_j is a monotonic transformation of θ (Lord, 1980, p.71),

$$\sigma^2(\hat{T}_j \mid \theta) = \sigma^2(\hat{T}_j \mid T_j) \approx I(T_j, \hat{T}_j)^{-1} \tag{15}$$

where

$I(T_j, \hat{T}_j)$ is the information that $\hat{T}_j$ contributes about T_j.

Given these results, Lord (1980, p. 79 and 85) produces

$$I(T_j, \hat{T}_j) = \frac{I(\theta, \hat{T}_j)}{\left(\partial T_j / \partial \theta\right)^2}, \tag{16}$$

and

$$I(\theta, \hat{T}_j) \approx I(\theta, \hat{\theta}). \tag{17}$$

Thus,

$$\sigma^2(\hat{T}_j \mid \theta) \approx \frac{\left[\frac{1}{n_j} \sum_{i=1}^{k_j} \hat{T}_{ij}(\theta)\right]^2}{I(\theta, \hat{\theta})}$$

and the parameters of the prior beta distribution for T_j can be expressed in terms of the parameters of the 3PL IRT and 2PPC models. Furthermore, the parameters of the posterior distribution of T_j also can be expressed in terms of the IRT parameters:

$$p_j = \hat{T}_j n_j^* + x_j, \tag{18}$$

and

$$q_j = \left[1 - \hat{T}_j\right] n_j^* + n_j - x_j. \tag{19}$$

The SPI is

$$\tilde{T}_j = \frac{p_j}{p_j + q_j} \tag{20}$$

$$= \frac{\hat{T}_j n_j^* + x_j}{n^*{}_j + n_j}. \tag{21}$$

The SPI can also be written in terms of the relative contribution of the prior estimate $\hat{T}_j$ and the observed proportion of maximum raw (correct score) (OPM), x_j / n_j, as

$$\tilde{T}_j = w_j \hat{T}_j + (1 - w_j)\left[x_j / n_j\right]. \tag{22}$$

w_j, a function of the mean and variance of the prior distribution, is the relative weight given to the prior estimate:

$$w_j = \frac{n_j^*}{n_j^* + n_j}. \tag{23}$$

The term n_j^* may be interpreted as the contribution of the prior in terms of theoretical numbers of items.

<u>Check on Consistency and Adjustment of Weight Given to Prior Estimate</u>

The item responses are assumed to be described by $P_i(\hat{\theta})$ or $P_{im}(\hat{\theta})$, depending on the type of item. Even if the IRT model accurately described item performance over examinees, their item responses grouped by strand may be multidimensional. For example, a particular examinee may be able to perform difficult addition but not easy subtraction. Under these circumstances, it is not appropriate to pool the prior estimate, $\hat{T}_j$, with x_j / n_j. In calculating the SPI, the following statistic was used to identify examinees with unexpected performance on the strands in a test:

$$Q = \sum_{j=1}^{J} n_j \left(\frac{x_j}{n_j} - \hat{T}_j\right)^2 / (\hat{T}_j(1 - \hat{T}_j)). \tag{24}$$

If $Q \le \chi^2(J, .10)$, the weight, w_j, is computed and the SPI is produced. If $Q > \chi^2(J, .10)$, n_j^* and subsequently w_j is set equal to 0 and the OPM is used as the estimate of strand performance.

As previously noted, the prior is estimated using an ability estimate based on responses to all the items (including the items of strand *j*) and hence is not independent of X_j. An adjustment for the overlapping information that requires minimal computation is to multiply the test information in equation 5 by the factor $(n - n_j) / n$. The application of this factor produces an “adjusted” SPI estimate that can be compared to the “unadjusted” estimate.

Possible Violations of the Assumptions

Even if the IRT model fits the test items, the responses for a given examinee, grouped by strand, may be multidimensional. In these cases, it would not be appropriate to pool the prior estimate, $\hat{T}_j$, with x_j / n_j. A chi-square fit statistic is used to evaluate the observed proportion of maximum raw score (OPM) relative to that predicted for the items in the strand on the basis of the student's overall trait estimate. If the chi-square is significant, the prior estimate is not used and the OPM obtained becomes the student's strand score.

If the items in the strand do not permit guessing, it is reasonable to assume $\hat{T}_j$, the expected proportion correct of the maximum score for a strand, will be greater or equal to zero. If correct guessing is possible, as it is with MC items, there will be a non-zero lower limit to $\hat{T}_j$, and a three-parameter beta distribution, in which $\hat{T}_j$ is greater than or equal to this lower limit (Johnson & Kotz, 1979, p. 37), would be more appropriate. The use of the two-parameter beta distribution would tend to underestimate T_j among very low-performing examinees. While working with tests containing exclusively MC items, Yen found that there does not appear to be a practical importance to this underestimation (Yen, 1997). The impact of any such effect would be reduced as the proportion of CR items in the test increases. The size of this effect, nonetheless, was evaluated using simulations (Yen, Sykes, Ito, & Julian, 1997).

The SPI procedure assumes that $p(X_j T_j)$ is a binomial distribution. This assumption is appropriate only when all the items in a strand have the same Bernoulli item response function. Not only do real items differ in difficulty, but when there are mixed-item types, X_j is not the sum of n_j independent Bernoulli variables. It is instead the total raw score. In essence, the simplifying assumption has been made that each CR item with a maximum score of $1_j - 1$ is the sum of $1_j - 1$ independent Bernoulli variables. Thus, a complex compound distribution is theoretically more applicable than the binomial. Given the complexity of working with such a model, it appears valuable to determine if the simpler model described here is sufficiently accurate to be useful.

Finally, because the prior estimate of $T_j, \hat{T}_j$, is based on performance on the entire test, including strand *j*, the prior estimate is not independent of X_j. The smaller the ratio n_j / n, the less impact this dependence will have. The effect of the overlapping information would be to understate the width of the credibility interval. The extent to which the size of the credibility interval is too small was examined (Yen et al, 1997) by simulating strands that contained varying proportions of the total test points.

References

CTB/McGraw-Hill. (2000). *TerraNova* 2nd Edition. Monterey, CA.

Fitzpatrick, A. R., V. Link, W. M. Yen, G. Burket, K. Ito & R. Sykes (1996). Scaling performance assessments: A comparison between one-parameter and two-parameter partial credit models. *Journal of Educational Measurement,* 33, 291–314.

Johnson, N. L. & S. Kotz (1970). *Distributions in statistics: Continuous univariate distributions* (Vol. 2). New York: John Wiley.

Lord, F. M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement,* 16, 159–176.

Novick, M. R. & P. H. Jackson (1974). *Statistical methods for educational and psychological research.* New York: McGraw-Hill.

Yen, W. M. (1997). The technical quality of performance assessments: Standard errors of percents of students reaching standards. *Educational Measurement: Issues and Practice*, 5–15.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement,* 30, 187–213.

Yen, W. M., R. C. Sykes, K. Ito & M. Julian (1997). *A Bayesian/IRT index of objective performance for tests with mixed-item types.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.