

# Chapter 3

## Scoring and Scaling Educational Tests

Michael J. Kolen, Ye Tong, and Robert L. Brennan

The numbers that are associated with examinee performance on educational or psychological tests are defined through the process of *scaling*. This process produces a *score scale*, and the scores that are reported to examinees are referred to as *scale scores*. Kolen (2006) referred to the term *primary score scale*, which is the focus of this chapter, as the scale that is used to underlie psychometric properties for tests.

A key component in the process of developing a score scale is the *raw score* for an examinee on a test, which is a function of the *item scores* for that examinee. Raw scores can be as simple as a sum of the item scores or be so complicated that they depend on the entire pattern of item responses.

Raw scores are transformed to scale scores to facilitate the meaning of scores for test users. For example, raw scores might be transformed to scale scores so that they have predefined distributional properties for a particular group of examinees, referred to as a *norm group*. Normative information might be incorporated by constructing scale scores to be approximately normally distributed with a mean of 50 and a standard deviation of 10 for a national population of examinees. In addition, procedures can be used for incorporating content and score precision information into score scales.

The purpose of this chapter is to describe methods for developing score scales for educational and psychological tests. Different types of raw scores are considered along with models for transforming raw scores to scale scores. Both traditional and item response theory (IRT) methods are considered. The focus of this chapter is

---

M.J. Kolen (✉)

University of Iowa, 224 B1 Lindquist Center, Iowa City, IA 52242, USA  
e-mail: michael-kolen@uiowa.edu

Y. Tong

Pearson, 2510 North Dodge Street, Iowa City, IA 52245, USA  
e-mail: ye.tong@pearson.com

R.L. Brennan

University of Iowa, 210D Lindquist Center, Iowa City, IA 52242, USA  
e-mail: robert-brennan@uiowa.edu

on fixed tests, rather than on computer-adaptive tests (Drasgow, Luecht, & Bennett, 2006), although many of the issues considered apply to both.

### 3.1 Unit and Item Scores

Kolen (2006) distinguished unit scores from item scores. A *unit score* is the score on the smallest unit on which a score is found, which is referred to as a *scoreable unit*. An item score is a score over all scoreable units for an item.

For multiple-choice test questions that are scored right-wrong, unit scores and item scores often are the same. Such scores are either incorrect (0) or correct (1). Unit and item scores are often distinguishable when judges score the item responses. As an example, consider an essay item that is scored 1 (*low*) through 5 (*high*) by each of two judges, with the item score being the sum of scores over the two judges. In this situation, there is a unit score for Judge 1 (range 1 to 5), a unit score for Judge 2 (range 1 to 5), and an item score over the two judges (range 2 to 10).

Or, consider a situation in which a block of five questions is associated with a reading passage. If a test developer is using IRT and is concerned that there might be conditional dependencies among responses to questions associated with a reading passage, the developer might treat the questions associated with passage as a single item, with scores on this item being the number of questions associated with the passage that the examinee answers correctly. In this case, each question would have a unit score of 0 or 1, and the item score would range from 0 to 5. According to Kolen (2006), “The characteristic that most readily distinguishes unit scores from item scores is... whereas there may be operational dependencies among unit scores, item scores are considered operationally independent” (p. 157).

Let  $V_i$  be a random variable indicating score on item  $i$  and  $v_i$  be a particular score. For a dichotomously scored item,  $v_i = 0$  when an examinee incorrectly answers the item and  $v_i = 1$  when an examinee correctly answers the item.

Consider the essay item described earlier. For this item,  $V_i = 2, 3, \dots, 10$  represent the possible scores for this item. In this chapter, it is assumed that when polytomously scored items are used, they are ordered response item scores represented by consecutive integers. Higher item scores represent more proficient performance on the item.

Item types exist where responses are not necessarily ordered, such as nominal response scoring. Such item types are not considered in this chapter.

### 3.2 Traditional Raw Scores

Let  $X$  refer to the raw score on a test. The *summed score*,  $X$ , is defined as

$$X = \sum_{i=1}^n V_i, \quad (3.1)$$

where  $n$  is the number of items on the test. Equation 3.1 is often used as a raw score when all of the items on a test are of the same format.

The *weighted summed score*,  $X_w$ ,

$$X_w = \sum_{i=1}^n w_i V_i, \quad (3.2)$$

uses weights,  $w_i$ , to weight the item score for each item. Various procedures for choosing weights include choosing weights that maximize score reliability and choosing weights so that each item contributes a desired amount to the raw score.

### 3.3 Traditional Scale Scores

Raw scores such as those in Equations 3.1 and 3.2 have limitations as primary score scales for tests. *Alternate test forms* are test forms that are built to a common set of content and statistical specifications. With alternate test forms, the raw scores typically do not have a consistent meaning across forms. For this reason, scores other than raw scores are used as primary score scales, whenever alternate forms of a test exist. The primary score scale typically is developed with an initial form of the test, and test equating methods (Holland & Dorans, 2006; Kolen & Brennan, 2004) are used to link raw scores on new forms to the score scale.

The raw score is transformed to a scale score. For summed scores, the scale score,  $S_X$ , is a function of the summed score,  $X$ , such that  $S_X = S_X(X)$ . This transformation often is provided in tabular form. For weighted summed scores, the scale score,  $S_{wX}$ , is a function of the weighted summed score,  $X_w$ , such that  $S_{wX} = S_{wX}(X_w)$ . For weighted summed scores, there are often many possible scale scores, so a continuous function may be used. In either case, scale scores are typically rounded to integers for score reporting purposes.

Linear or nonlinear transformations of raw scores are used to produce scale scores that can be meaningfully interpreted. Normative, score precision, and content information can be incorporated. Transformations that can be used to incorporate each of these types of meaning are considered next.

#### 3.3.1 Incorporating Normative Information

Incorporating normative information begins with the administration of the test to a norm group. Statistical characteristics of the scale score distribution are set relative to this norm group. The scale scores are meaningful to the extent that the norm group is central to score interpretation.

For example, a third-grade reading test might be administered to a national norm group intended to be representative of third graders in the nation. The mean and standard deviation of scale scores on the test might be set to particular values for

this norm group. By knowing the mean and standard deviation of scale scores, test users would be able to quickly ascertain, for example, whether a particular student's score was above the mean. This information would be relevant to the extent that scores for the norm group are central to score interpretation. Kolen (2006, pp. 163–164) provided equations for linearly transforming raw scores to scale scores with a particular mean and standard deviation.

Nonlinear transformations also are used to develop score scales. Normalized scores involve one such transformation. To normalize scores, percentile ranks of raw scores are found and then transformed using an inverse normal transformation. These normalized scores are then transformed to have a desired mean and standard deviation. Normalized scale scores can be used to quickly ascertain the percentile rank of a particular student's score using facts about the normal distribution. For example, with normalized scores, a score that is one standard deviation above the mean has a percentile rank of approximately 84. Kolen (2006, pp. 164–165) provided a detailed description of the process of score normalization.

Scale scores typically are reported to examinees as integer scores. For example, McCall (1939) suggested using T scores, which are scale scores that are normalized with an approximate mean of 50 and standard deviation of 10, with the scores rounded to integers. Intelligence test scores typically are normalized scores with a mean of 100 and a standard deviation of 15 or 16 in a national norm group (Angoff, 1971/1984, p. 525–526), with the scores rounded to integers.

### 3.3.2 *Incorporating Score Precision Information*

According to Flanagan (1951), scale score units should “be of an order of magnitude most appropriate to express their accuracy of measurement” (p. 246). Flanagan indicated that the use of too few score points fails to “preserve all of the information contained in raw scores” (p. 247). However, the use of too many scale score points might lead test users to attach significance to scale score differences that are predominantly due to measurement error.

Based on these considerations, rules of thumb have been developed to help choose the number of distinct score points to use for a scale. For example, the scale for the Iowa Tests of Educational Development (ITED, 1958) was constructed so that an approximate 50% confidence interval for true scores could be found by adding 1 scale score point to and subtracting 1 scale score point from an examinee's scale score. Similarly, Truman L. Kelley (W. H. Angoff, personal communication, February 17, 1987) suggested constructing scale scores so that an approximate 68% confidence interval could be constructed by adding 3 scale score points to and subtracting 3 scale score points from each examinee's scale score.

Kolen and Brennan (2004, pp. 346–347) showed that by making suitable assumptions, the approximate range of scale scores that produces the desired score scale property is

$$6 \frac{h}{z_\gamma \sqrt{1 - \rho_{XX'}}}, \quad (3.3)$$

where  $h$  is the width of the desired confidence interval (1 for the ITED rule, 3 for the Kelley rule),  $z_\gamma$  is the unit-normal score associated with the confidence coefficient  $\gamma$  ( $z_\gamma \approx .6745$  for the ITED rule and  $z_\gamma \approx 1$  for the Kelley rule), and  $\rho_{XX'}$  is test reliability. The result from Equation 3.3 is rounded to an integer. As an example, assume that test reliability is .91. Then for the ITED rule, Equation 3.3 indicates that 30 distinct scale score points should be used, and Kelley's rule indicates that 60 distinct score points should be used.

Noting that conditional measurement error variability is typically unequal along the score scale, Kolen (1988) suggested using a variance stabilizing transformation to equalize error variability. Kolen (1988) argued that when scores are transformed in this way, a single standard error of measurement could be used when reporting measurement error variability. He used the following arcsine transformation suggested by Freeman and Tukey (1950):

$$g(X) = .5 \left\{ \sin^{-1} \left[ \left( \frac{X}{n+1} \right) \right]^{\frac{1}{2}} + \sin^{-1} \left[ \left( \frac{X+1}{n+1} \right) \right]^{\frac{1}{2}} \right\}. \quad (3.4)$$

Scores transformed using Equation 3.4 are then transformed to have a desired mean and standard deviation and to have a reasonable number of distinct integer score points. Kolen, Hanson, and Brennan (1992) found that this transformation adequately stabilized error variance for tests with dichotomously scored items. Ban and Lee (2007) found a similar property for tests with polytomously scored items.

### 3.3.3 Incorporating Content Information

Ebel (1962) stated, "To be meaningful any test scores must be related to test content as well as to the scores of other examinees" (p. 18). Recently, focus has been on providing content meaningful scale scores.

One such procedure, *item mapping*, was reviewed by Zwirk, Senturk, Wang, and Loomis (2001). In item mapping, test items are associated with various scale score points. For dichotomously scored items, the probability of correct response on each item is regressed on scale score. The *response probability* (RP) level is defined as the probability (expressed as a percentage) of correct response on a test-item given scale score that is associated with mastery, proficiency, or some other category as defined by the test developer. The same RP level is used for all dichotomously scored items on the test. Using regressions of item score on scale score, an item is said to map at the scale score associated with an RP of correctly answering the item. RP values typically range from .5 to .8. Additional criteria are often used when choosing items to report on an item map, such as item discrimination and test developer judgment. Modifications of the procedures are used with polytomously scored items. The outcome of an item mapping procedure is a map illustrating which items correspond to each of an ordered set of scale scores.

Another way to incorporate content information is to use *scale anchoring*. The first step in scale anchoring is to develop an item map. Then, a set of scale score points is chosen, such as a selected set of percentiles. Subject-matter experts review the items that map near each of the selected points and develop general statements that represent the skills of the examinees scoring at each point. See Allen, Carlson, and Zelenak (1999) for an example of scale anchoring with the National Assessment of Educational Progress and ACT (2001) for an example of scale anchoring as used with the ACT Standards for Transition.

*Standard setting* procedures, as recently reviewed by Hambleton and Pitoniak (2006), begin with a statement about what competent examinees know and are able to do. Structured judgmental processes are used to find the scale score point that differentiates candidates who are minimally competent from those who are less than minimally competent. In achievement testing situations, various achievement levels are often stated, such as basic, proficient, and advanced. Judgmental standard-setting techniques are used to find the scale score points that differentiate between adjacent levels.

### 3.3.4 Using Equating to Maintain Score Scales

Equating methods (Holland & Dorans, 2006; Kolen & Brennan, 2004) are used to maintain scale scores as new forms are developed. For equating to be possible, the new forms must be developed to the same content and statistical specifications as the form used for scale construction. With traditional scaling and equating methodology, a major goal is to transform raw scores to scale scores on new test forms so that the distribution of scale scores is the same in a population of examinees.

## 3.4 IRT Proficiency Estimates

Traditional methods focus on scores that are observed rather than on true scores or IRT proficiencies. Procedures for using psychometric methods to help evaluate scale scores with traditional methods exist and are described in a later section of this chapter. First, scale scores based on IRT (Thissen & Wainer, 2001) are considered.

The development of IRT scaling methods depends on the use of an IRT model. In this section, IRT models are considered in which examinee proficiency,  $\theta$ , is assumed to be unidimensional. A *local independence assumption* also is required, in which, conditional on proficiency, examinee responses are assumed to be independent. The focus of the IRT methods in this section is on polytomously scored

items. Note, however, that dichotomously scored items can be viewed as polytomously scored items with two response categories (wrong and right).

A curve is fit for each possible response to an item that relates probability of that response given proficiency that is symbolized as  $P(V_i = v_i|\theta)$  and is referred to as the *category response function*. IRT models considered here have the responses ordered a priori, where responses associated with higher scores are indicative of greater proficiency. Popular models for tests containing dichotomously scored items are the Rasch, two-parameter logistic, and three-parameter logistic models. Popular models for tests containing polytomously scored items are the graded-response model, partial-credit model, and generalized partial-credit model. Nonparametric models also exist. See Yen and Fitzpatrick (2006) and van der Linden and Hambleton (1997) for reviews of many of these models.

In IRT, the category response functions are estimated for each item. Then, proficiency is estimated for each examinee. In this chapter, initial focus is on IRT proficiency estimation. Later, the focus is on the transformed (often linearly, but sometimes nonlinearly) proficiencies typically used when developing scale scores. For this reason, IRT proficiency estimates can be thought of as raw scores that are subsequently transformed to scale scores.

Estimates of IRT proficiency can be based on summed scores (Equation 3.1), weighted summed scores (Equation 3.2), or on *complicated scoring functions* that can be symbolized as

$$X_c = f(V_1, V_2, \dots, V_n), \quad (3.5)$$

where  $f$  is the function used to convert item scores to total score. Some models, such as the Rasch model, consider only summed scores. With other models, the psychometrician can choose which scoring function to use. Procedures for estimating IRT proficiency are described next.

### 3.4.1 IRT Maximum Likelihood Scoring

IRT maximum likelihood scoring requires the use of a complicated scoring function for many IRT models. Under the assumption of local independence,  $\theta$  is found that maximizes the likelihood equation,

$$L = \prod_{i=1}^n p(V_i = v_i|\theta), \quad (3.6)$$

and it is symbolized as  $\hat{\theta}_{MLE}$ .

### 3.4.2 IRT With Summed Scores Using the Test Characteristic Function

In IRT it is possible to estimate proficiency as a function of summed scores or weighted summed scores. Assume that item  $i$  is scored in  $m_i$  ordered categories, where the categories are indexed  $k = 1, 2, \dots, m_i$ . Defining  $W_{ik}$  as the score associated with item  $i$  and category  $k$ , the *item response function* for item  $i$  is defined as

$$\tau_i(\theta) = \sum_{k=1}^{m_i} W_{ik} \cdot P(V_i = k|\theta), \quad (3.7)$$

which represents the expected score on item  $i$  for an examinee of proficiency  $\theta$ . For IRT models with ordered responses, it is assumed that  $\tau_i(\theta)$  is monotonic increasing.

The *test characteristic function* is defined as the sum, over test items, of the item response functions such that

$$\tau(\theta) = \sum_{i=1}^n \tau_i(\theta), \quad (3.8)$$

which represents the true score for an examinee of proficiency  $\theta$ . This function is also monotonic increasing.

A weighted test characteristic function also can be defined as

$$\tau_w(\theta) = \sum_{i=1}^n w_i \tau_i(\theta), \quad (3.9)$$

where the  $w_i$  are positive-valued weights that are applied to each of the items when forming a total score.

An estimate of proficiency, based on a summed score for an examinee, can be found by substituting the summed score for  $\tau(\theta)$  in Equation 3.8 and then solving for  $\theta$  using numerical methods. Similarly, proficiency can be estimated for weighted sum scores. The resulting estimate using Equation 3.8 is referred to as  $\hat{\theta}_{TCF}$  and is monotonically related to the summed score. The resulting estimate using Equation 3.9 is referred to as  $\hat{\theta}_{wTCF}$  and is monotonically related to the weighted summed score.

### 3.4.3 IRT Bayesian Scoring With Complicated Scoring Functions

IRT Bayesian estimates of proficiency can make use of a complicated scoring function. In addition, they require specification of the distribution of proficiency in the population,  $g(\theta)$ . The Bayesian modal estimator is the  $\theta$  that maximizes



$$L \cdot g(\theta) = \prod_{i=1}^n P(V_i = v_i | \theta) \cdot g(\theta) \quad (3.10)$$

and is symbolized as  $\hat{\theta}_{BME}$ . The Bayesian expected a posteriori (EAP) estimator is the mean of the posterior distribution and is calculated as

$$\begin{aligned} \hat{\theta}_{EAP} &= E(\theta | V_1 = v_1, V_2 = v_2, \dots, V_n = v_n) \\ &= \frac{\int \theta \prod_{i=1}^n P(V_i = v_i | \theta) g(\theta) d\theta}{\int \prod_{i=1}^n P(V_i = v_i | \theta) g(\theta) d\theta} \end{aligned} \quad (3.11)$$

### 3.4.4 Bayesian Scoring Using Summed Scores

A Bayesian EAP estimate of proficiency based on the summed score is

$$\begin{aligned} \hat{\theta}_{sEAP} &= E(\theta | X) \\ &= \frac{\int \theta \cdot P(X = x | \theta) \cdot g(\theta) d\theta}{\int P(X = x | \theta) \cdot g(\theta) d\theta} \end{aligned} \quad (3.12)$$

The term  $P(X = x | \theta)$  represents the probability of earning a particular summed score given proficiency and can be calculated from item parameter estimates using a recursive algorithm provided by Thissen, Pommerich, Billeaud, and Williams (1995) and illustrated by Kolen and Brennan (2004, pp. 219-221), which is a generalization of a recursive algorithm developed by Lord and Wingersky (1984).

Concerned that the estimate in Equation 3.12 treats score points on different item types as being equal, Rosa, Swygert, Nelson, and Thissen (2001) presented an alternative Bayesian EAP estimator. For this alternative, define  $X_1$  as the summed score on the first item type and  $X_2$  as summed score on the second item type. The EAP is the expected proficiency given scores on each item type and is

$$\begin{aligned} \hat{\theta}_{s2EAP} &= E(\theta | X_1, X_2) \\ &= \frac{\int \theta \cdot P(X_1 = x_1 | \theta) \cdot P(X_2 = x_2 | \theta) \cdot g(\theta) d\theta}{\int P(X_1 = x_1 | \theta) \cdot P(X_2 = x_2 | \theta) \cdot g(\theta) d\theta} \end{aligned} \quad (3.13)$$

where  $P(X_1 = x_1|\theta)$  and  $P(X_2 = x_2|\theta)$  are calculated using the recursive algorithm provided by Thissen et al. (1995). Note that this estimate is, in general, different for examinees with different combinations of scores on the two item types. Rosa et al. (2001) presented results for this method in a two-dimensional scoring table, with summed scores on one item type represented by the rows and summed scores on the other item type represented by the columns. Rosa et al. indicated that this method can be generalized to tests with more than two item types. Thissen, Nelson, and Swygert (2001) provided an approximate method in which the EAP is estimated separately for each item type and then a weighted average is formed. Bayesian EAP estimates have yet to be developed based on the weighted summed scores defined in Equation 3.2.

### 3.4.5 Statistical Properties of Estimates of IRT Proficiency

The maximum likelihood estimator  $\hat{\theta}_{MLE}$  and test characteristic function estimators  $\hat{\theta}_{TCF}$  and  $\hat{\theta}_{wTCF}$  do not depend on the distribution of proficiency in the population,  $g(\theta)$ . All of the Bayesian estimators depend on  $g(\theta)$ .

$\hat{\theta}_{MLE}$ ,  $\hat{\theta}_{TCF}$ , and  $\hat{\theta}_{wTCF}$  do not exist (are infinite) for examinees whose item score is the lowest possible score on all of the items. In addition, these estimators do not exist for examinees whose item score is the highest possible score on all of the items. Other extreme response patterns exist for which  $\hat{\theta}_{MLE}$  does not exist. Also, for models with a lower asymptote item parameter, like the three-parameter logistic model,  $\hat{\theta}_{TCF}$  does not exist for summed scores that are below the sum, over items, of the lower asymptote parameters. A similar issue is of concern for  $\hat{\theta}_{wTCF}$ . In practice, ad hoc rules are used to assign proficiency estimates for these response patterns or summed scores. The Bayesian estimators typically exist in these situations, which is a benefit of these estimators.

The maximum likelihood estimator of proficiency,  $\hat{\theta}_{MLE}$ , is consistent (Lord, 1980, p. 59), meaning that it converges to  $\theta$  as the number of items becomes large. Thus,

$$E(\hat{\theta}_{MLE}|\theta) \approx \theta. \quad (3.14)$$

Note also that  $E(X|\theta) = \tau(\theta) = \sum_{i=1}^n \tau_i(\theta)$ , which means that the summed score  $X$  is an unbiased estimate of true summed score  $\tau$ . This suggests that  $E(\hat{\theta}_{TCF}|\theta)$  is close to  $\theta$ .

The Bayesian estimators are shrinkage estimators intended to be biased when a test is less than perfectly reliable. So for most values of  $\theta$ ,

$$E(\hat{\theta}_{EAP}|\theta) \neq \theta. \quad (3.15)$$

Defining  $\mu_\theta$  as the mean of the distribution of proficiency,

$$\begin{aligned} \text{If } \theta < \mu_\theta, \text{ then } E(\hat{\theta}_{EAP}|\theta) &> \theta \\ \text{If } \theta > \mu_\theta, \text{ then } E(\hat{\theta}_{EAP}|\theta) &< \theta. \end{aligned} \quad (3.16)$$

Similar relationships hold for the other Bayesian estimators.

Test information is a central concept in IRT when considering conditional error variability in estimating IRT proficiency. Conditional error variance in estimating proficiency in IRT using maximum likelihood is equal to 1 divided by test information. Expressions for conditional error variances of the maximum likelihood estimators,  $\text{var}(\hat{\theta}_{MLE}|\theta)$ , and for the test characteristic function estimators,  $\text{var}(\hat{\theta}_{TCF}|\theta)$  and  $\text{var}(\hat{\theta}_{wTCF}|\theta)$ , for dichotomous models have been provided by Lord (1980) and for polytomous models by Muraki (1993), Samejima (1969), and Yen and Fitzpatrick (2006). Note that the square root of the conditional error variance is the *conditional standard error of measurement* for estimating IRT proficiency.

An expression for the conditional error variance for Bayesian EAP estimators was provided by Thissen and Orlando (2001) and is as follows for  $\hat{\theta}_{EAP}$ :

$$\begin{aligned} \text{var}(\hat{\theta}_{EAP}|V_1 = v_1, V_2 = v_2, \dots, V_n = v_n) \\ = \frac{\int (\hat{\theta}_{EAP} - \theta)^2 \prod_{i=1}^n P(V_i = v_i|\theta)g(\theta)d\theta}{\int \prod_{i=1}^n P(V_i = v_i|\theta)g(\theta)d\theta}. \end{aligned} \quad (3.17)$$

Similar expressions can be used for  $\hat{\theta}_{sEAP}$ .

Note that the Bayesian conditional variances are conditional on examinee response patterns, which is typical for Bayesian estimators, rather than on  $\theta$ , as is the case with the maximum likelihood and test characteristic function estimators. This observation highlights a crucial difference in the meaning of conditional error variances for Bayesian and maximum likelihood estimates of proficiency.

The following relationship is expected to hold:

$$\text{var}(\hat{\theta}_{TCF}|\theta) \geq \text{var}(\hat{\theta}_{MLE}|\theta) \geq \text{var}(\hat{\theta}_{EAP}|\theta). \quad (3.18)$$

Note that  $\text{var}(\hat{\theta}_{TCF}|\theta) \geq \text{var}(\hat{\theta}_{MLE}|\theta)$  because  $\hat{\theta}_{TCF}$  is based on summed scores, which leads to a loss of information as compared to  $\hat{\theta}_{MLE}$ . Also,  $\text{var}(\hat{\theta}_{MLE}|\theta) \geq \text{var}(\hat{\theta}_{EAP}|\theta)$ , because Bayesian estimators are shrinkage estimators that generally have smaller error variances than maximum likelihood estimators. However, the Bayesian estimators are biased, which could cause the conditional mean-squared error for  $\hat{\theta}_{EAP}$ , defined as  $MSE(\hat{\theta}_{EAP}|\theta) = E[(\hat{\theta}_{EAP} - \theta)|\theta]^2$ , to be greater than the mean-squared error for  $\hat{\theta}_{MLE}$ . Note that conditional mean-squared error takes into account both error variance and bias. In addition it is expected that

$$\text{var}(\hat{\theta}_{EAP}|\theta) \geq \text{var}(\hat{\theta}_{sEAP}|\theta), \tag{3.19}$$

because there is less error involved with pattern scores than summed scores, resulting in less shrinkage with  $\hat{\theta}_{EAP}$  than with  $\hat{\theta}_{sEAP}$ .

The relationships between conditional variances have implications for the marginal variances. In particular, the following relationship is expected to hold if the distribution of  $\theta$  is well specified:

$$\text{var}(\hat{\theta}_{TCF}) \geq \text{var}(\hat{\theta}_{MLE}) \geq \text{var}(\hat{\theta}_{EAP}) \geq (\hat{\theta}_{sEAP}). \tag{3.20}$$

As illustrated in the next section, the inequalities can have practical implications.

### 3.4.6 Example: Effects of Different Marginal Distributions on Percentage Proficient

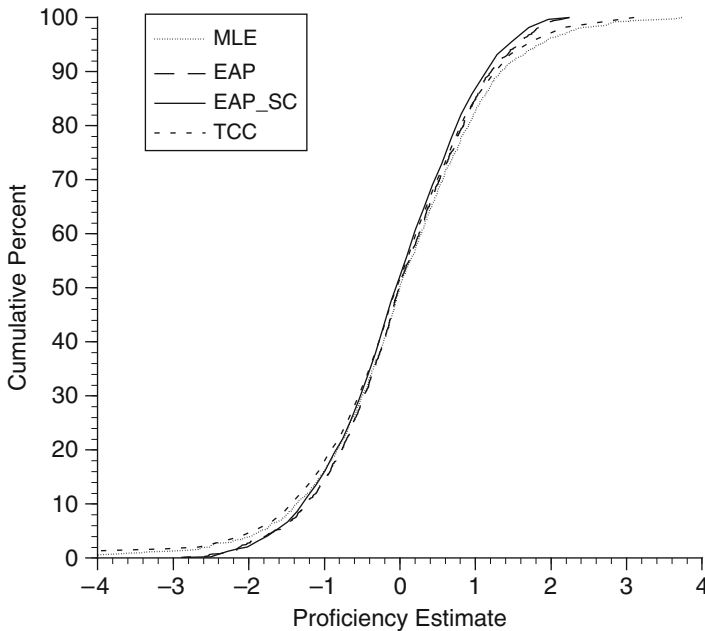
Suppose there are four performance levels (Levels I through IV) for a state assessment program. Based on a standard setting study, cut scores on the  $\theta$  scale are  $-0.8$  for Level I-II,  $0.2$  for Level II-III and  $1.3$  for Level III-IV. As illustrated in this section, the choice of proficiency estimator can have a substantial effect on the percentage of students classified at each of the performance levels.

In this hypothetical example, the scaling data for Grade 7 of the Vocabulary test of the Iowa Tests of Basic Skills were used. The test contains 41 multiple-choice items. For more information on the dataset used, see Tong and Kolen (2007). For each student in the dataset ( $N = 1,199$ ),  $\hat{\theta}_{MLE}$ ,  $\hat{\theta}_{EAP}$ ,  $\hat{\theta}_{sEAP}$ , and  $\hat{\theta}_{TCF}$  were computed based on the same set of item parameter estimates. Table 3.1 shows the mean and standard deviation (*SD*) of the proficiency estimates for all the students included in the example. Figure 3.1 shows the cumulative frequency distribution of the proficiency estimates for these students. The variabilities of these estimators are ordered, as expected based on Equation 3.20, as

$$SD(\hat{\theta}_{sEAP}) < SD(\hat{\theta}_{EAP}) < SD(\hat{\theta}_{MLE}) < SD(\hat{\theta}_{TCF}).$$

**Table 3.1** Example: Effects of IRT Proficiency Estimator on Percent in Proficiency Level

Proficiency estimator	<i>M</i>	<i>SD</i>	Percentage proficiency by level			
			I	II	III	IV
$\hat{\theta}_{MLE}$	0.012	1.143	20.77	35.95	32.53	10.76
$\hat{\theta}_{EAP}$	-0.002	0.949	19.27	38.70	33.86	8.17
$\hat{\theta}_{sEAP}$	0.000	0.933	19.43	36.53	37.20	6.84
$\hat{\theta}_{TCF}$	-0.003	1.164	22.02	33.94	33.53	10.51



**Fig 3.1** Cumulative distributions for various proficiency estimators

From Figure 3.1, the cumulative distributions for  $\hat{\theta}_{EAP}$  and  $\hat{\theta}_{SEAP}$  are similar to one other and the cumulative distributions for  $\hat{\theta}_{TCF}$  and  $\hat{\theta}_{MLE}$  are similar to one other.

Using the  $\theta$  cut scores from standard setting, students were classified into each of the four performance levels using the four proficiency estimates. The percentage in each level for each of the estimators is reported in Table 3.1. As can be observed,  $\hat{\theta}_{MLE}$  and  $\hat{\theta}_{TCF}$  tend to produce larger percentages of students in Levels I and IV, consistent with the observation that these estimators have relatively larger variability. Of the 1,199 students in the data, 31 students (about 13%) had different performance-level classifications using different proficiency estimators. The differences were within one performance level. These results illustrate that, in practice, the choice of IRT proficiency estimator can affect the proficiency level reported for a student.

### 3.5 IRT Scale Scores

IRT scale scores often are developed by linearly transforming IRT proficiencies and then rounding to integers. When a linear transformation is used, the estimators and their statistical properties can be found directly based on the linear transformation. Sometimes IRT scale scores are transformed nonlinearly to scale scores. Lord (1980, pp. 84–88) argued that a nonlinear transformation of  $\theta$  could be preferable to  $\theta$ .

Define  $h$  as a continuous monotonically increasing function of proficiency, such that

$$\theta^* = h(\theta). \quad (3.21)$$

Lord (1980, pp. 187–188) showed that the maximum likelihood estimator for a nonlinear transformed proficiency could be found by applying the nonlinear transformation to the maximum likelihood estimator parameter estimate. That is, he showed that

$$\hat{\theta}^*_{MLE} = h(\hat{\theta}_{MLE}). \quad (3.22)$$

Estimates based on the test characteristic function are found by a similar substitution. Lord (1980, pp. 187–188) also showed that Bayesian estimators do not possess this property. Thus, for example,

$$\hat{\theta}^*_{EAP} \neq h(\hat{\theta}_{EAP}). \quad (3.23)$$

To find  $\hat{\theta}^*_{EAP}$  would require computing the estimate using Equation 3.11 after substituting  $\theta^*$  for each occurrence of  $\theta$ .

One nonlinear transformation that is often used is the *domain score* (Bock, Thissen, & Zimowski, 1997; Pommerich, Nicewander, & Hanson, 1999), which is calculated as follows:

$$\theta^*_{domain} = \frac{1}{n_{domain}} \sum_{i=1}^{n_{domain}} \tau_i(\theta), \quad (3.24)$$

where  $\tau_i(\theta)$  is defined in Equation 3.7, and the summation is over all  $n_{domain}$  items in the domain, where the domain is a large number of items intended to reflect the content that is being assessed. Substituting  $\hat{\theta}_{MLE}$  for  $\theta$  in Equation 3.24 produces the maximum likelihood estimator of  $\theta^*_{domain}$ . However, substituting  $\hat{\theta}_{EAP}$  for  $\theta$  in Equation 3.24 does *not* produce a Bayesian EAP estimator of  $\theta^*_{domain}$ .

### 3.6 Multidimensional IRT Raw Scores for Mixed Format Tests

In applying IRT with mixed item types, an initial decision that is made is whether or not a single dimension can be used to describe performance. Rodriguez (2003) reviewed the construct equivalence of multiple-choice and constructed-response items. He concluded that these item types typically measure different constructs, although in certain circumstances the constructs are very similar. Wainer and Thissen (1993) argued that the constructs often are similar enough that the mixed item types can be reasonably analyzed with a unidimensional model. If a test developer decides that a multidimensional model is required, it is sometimes

possible to analyze each item type using a unidimensional model. IRT proficiency can be estimated separately for each item type, and then a weighted composite of the two proficiencies computed as an overall estimate of proficiency. This sort of procedure was used, for example, with the National Assessment of Educational Progress Science Assessment (Allen et al., 1999).

### 3.7 Psychometric Properties of Scale Scores

Psychometric properties of scale scores include (a) the expected (true) scale score, (b) the conditional error variance of scale scores, and (c) the reliability of scale scores for an examinee population. In addition, when alternate forms of a test exist, psychometric properties of interest include (a) the extent to which expected scale scores are the same on the alternate forms, often referred to as *first-order equity*; (b) the extent to which the conditional error variance of scale scores is the same on the alternate forms, often referred to as *second-order equity*; and (c) the extent to which reliability of scale scores is the same on alternate forms.

Assuming that scale scores are a function of summed scores of a test consisting of dichotomously scored items, Kolen et al. (1992) developed procedures for assessing these psychometric properties using a strong true-score model. For the same situation, Kolen, Zeng, and Hanson (1996) developed procedures for assessing these psychometric properties using an IRT model. Wang, Kolen, and Harris (2000) extended the IRT procedures to summed scores for polytomous IRT models.

The Wang et al. (2000) approach is used to express the psychometric properties as follows. Recall that  $S_X(X)$  represents the transformation of summed scores to scale scores. The expected (true) scale score given  $\theta$  is expressed as

$$\tau_{S_X} = \sum_{j=\min X}^{\max X} S_X(j) \cdot P(X = j|\theta), \quad (3.25)$$

where  $P(X = j|\theta)$  is calculated using a recursive algorithm (Thissen et al., 1995), and  $\min X$  and  $\max X$  are the minimum and maximum summed score. Conditional error variance of scale scores is expressed as

$$\text{var}(S_X|\theta) = \sum_{j=\min X}^{\max X} [S_X(j) - \tau_{S_X}]^2 \cdot P(X = j|\theta). \quad (3.26)$$

Reliability of scale scores is expressed as

$$\rho(S_X, S_{X'}) = 1 - \frac{\int \text{var}(S_X|\theta)g(\theta)d\theta}{\sigma_{S_X}^2}, \quad (3.27)$$

where  $\sigma_{S_x}^2$  is the variance of scale scores in the population. Using examples from operational testing programs, this framework has been used to study the relationship between  $\theta$  and true scale score, the pattern of conditional standard errors of measurement, the extent to which the arcsine transformation stabilizes error variance, first-order equity across alternate forms, second-order equity across alternate forms, reliability of scale scores, and the effects of rounding on reliability for different scales (Ban & Lee, 2007; Kolen et al., 1992, 1996; Tong & Kolen, 2005; Wang et al., 2000). These procedures have yet to be extended to weighted summed scores or to more complex scoring functions.

When the IRT proficiency scale is nonlinearly transformed as in Equation 3.21, based on Lord (1980, p. 85) the conditional error variance of  $\hat{\theta}_{MLE}^*$  is approximated as

$$\text{var}\left(\hat{\theta}_{MLE}^*|\theta\right) \approx \left(\frac{d\theta^*}{d\theta}\right)^2 \text{var}\left(\hat{\theta}_{MLE}|\theta\right), \quad (3.28)$$

where  $\left(\frac{d\theta^*}{d\theta}\right)^2$  is the squared first derivative of the transformation of  $\theta$  to  $\theta^*$ . A similar relationship holds for  $\hat{\theta}_{TCC}^*$  and  $\hat{\theta}_{wTCC}^*$ . Note that this conditional error variance does not take rounding into account. To find the conditional error variance for Bayesian estimators for transformed variables, in Equation 3.17  $\theta$  is replaced by  $\theta^*$  and  $\hat{\theta}_{EAP}$  is replaced by  $\hat{\theta}_{EAP}^*$ . For these procedures, the transformation of  $\theta$  to  $\theta^*$  must be monotonic increasing and continuous. When the transformation to scale scores is not continuous, such as when scale scores are rounded to integers, these procedures at best can provide an approximation to conditional error variance. In such cases, simulation procedures can be used to estimate bias and conditional error variance of scale scores.

### 3.8 Concluding Comments

Currently a variety of raw scores is used with educational tests that include summed scores, weighted summed scores, and various IRT proficiency estimates. We have demonstrated that the choice of raw score has practical implications for the psychometric properties of scores, including conditional measurement error, reliability, and score distributions. Raw scores are transformed to scale scores to enhance score interpretations. The transformations can be chosen so as to incorporate normative, content, and score precision properties.