

*Center for Advanced Studies in  
Measurement and Assessment*

*CASMA Research Report*

*Number 7*

**A Bootstrap Procedure for Estimating  
Decision Consistency for  
Single-Administration  
Complex Assessments\***

*Robert L. Brennan*

*Lei Wan<sup>†</sup>*

June 2004

---

\* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, April, 2004. The authors thank Won-Chan Lee, Michael Kane, Michael Kolen, and Ping Yin for helpful comments on a previous draft. Date of last revision: June 2, 2004.

<sup>†</sup>Robert L. Brennan is E. F. Lindquist Chair in Measurement and Testing and Director, Center for Advanced Studies in Measurement and Assessment (CASMA), 297 Lindquist North, College of Education, University of Iowa, Iowa City, IA 52242 (email: robert-brennan@uiowa.edu). Lei Wan is a research assistant in CASMA (email: lei-wan@uiowa.edu).

Center for Advanced Studies in  
Measurement and Assessment (CASMA)  
College of Education  
University of Iowa  
Iowa City, IA 52242  
Tel: 319-335-5439  
Fax: 319-384-0505  
Web: [www.uiowa.edu/~casma](http://www.uiowa.edu/~casma)

All rights reserved

## Contents

<b>1</b>	<b>An Overview of the Bootstrap</b>	<b>3</b>
<b>2</b>	<b>Decision Consistency and Boot-<i>i</i></b>	<b>4</b>
2.1	Indices . . . . .	4
2.2	Boot- <i>i</i> Steps . . . . .	6
2.3	Dichotomous Data and the Binomial Model . . . . .	7
2.4	Polytomous Data and the Multinomial Model . . . . .	7
2.5	Parametric Boot- <i>i</i> Procedure . . . . .	8
<b>3</b>	<b>Complex Assessments and Stratified Boot-<i>i</i></b>	<b>9</b>
3.1	A Synthetic Example . . . . .	10
3.2	A Semi-real Example . . . . .	11
<b>4</b>	<b>Discussion and Conclusions</b>	<b>14</b>
<b>5</b>	<b>References</b>	<b>16</b>
<b>A</b>	<b>Some Characteristics of boot-<i>i</i> Replications</b>	<b>17</b>
A.1	EMS Equations and Estimated Variance Components . . . . .	18
A.2	Bias-Correcting Adjustments . . . . .	18
A.3	Expected Variance of Boot- <i>i</i> Person Scores . . . . .	19
A.4	Covariance of Original-Data and Boot- <i>i</i> Person Scores . . . . .	19
A.5	Parallel Forms and the Boot- <i>i</i> Procedure . . . . .	20
A.5.1	Equal Means and Variances . . . . .	21
A.5.2	Covariance of Original and Boot- <i>i</i> Scores . . . . .	21
A.5.3	Impossibility of Satisfying all Conditions . . . . .	22
A.6	Boot- <i>i</i> Replications and Conditional SEMs . . . . .	22

**Abstract**

For a test that consists of dichotomously-scored items, several approaches have been reported in the literature for estimating decision consistency based on a single administration of a test. It is relatively straightforward to extend these approaches to a test that consists of polytomously-scored items with the same number of score categories for each item. Single-administration decision-consistency has not been studied much, however, for “complex” assessments—e.g., those that involve mixtures of different types of items, with or without scaling and/or equating of scores, etc. This paper considers the use of a particular type of bootstrap procedure for estimating various types of decision consistency for individuals and groups who take a single administration of a complex assessment. The bootstrap procedure is easy to implement, no matter how complex the assessment may be.

For a test that consists of dichotomously-scored items, several approaches have been reported in the literature for estimating decision consistency based on a single administration of a test. The approaches can be categorized into two types—those that make distributional-form assumptions about true scores and those that do not. For example, Huynh (1976) assumed a beta distribution for true scores and a binomial distribution for errors. He then used the beta-binomial (or negative hypergeometric) distribution to estimate decision consistency for a group of examinees. Subsequently, Hanson and Brennan (1990) extended Huynh’s approach by using the four-parameter beta distribution for true scores; they also considered the compound binomial distribution for errors. Lee, Hanson, and Brennan (2002) used the same models to consider classification consistency with multiple cut scores. By contrast, Subkoviak (1976) suggested an approach to estimating decision consistency based on the binomial model for errors, without any distributional-form assumptions for true scores. His approach essentially estimates decision consistency one examinee at a time and then averages over examinees. In a sense, this paper extends Subkoviak’s approach to complex assessments.

The methods discussed in the previous paragraph (and others that might be named) assume that a test consists of dichotomously-scored items. There is very little in the literature that relates explicitly to single-administration decision consistency for a test that consists of polytomously-scored items. However, the Breyer and Lewis (1994) and Livingston and Lewis (1995) procedures, which are discussed later, could be used. Also, of course, decision consistency can be estimated rather easily for any data if one is prepared to assume that observed scores have a bivariate normal distribution with a correlation equal to the reliability of the test. The problem, of course, is that the normality assumption is usually highly implausible.

Since there is a paucity of literature dealing with decision consistency for tests that consist of polytomously-scored items, it is not surprising that there is also little literature on decision consistency with complex assessments that consist of both dichotomous and polytomous items, and possibly other complicating factors. Consider the following not-too-hypothetical examples that are arranged in order of increasing complexity (more or less).

1. A test consists of 40 undifferentiated multiple-choice items and 10 constructed response items scored on a three point scale. The total score is an unweighted sum of the item scores.
2. A test consists of 40 undifferentiated multiple-choice items, 4 constructed response items scored on a three point scale, and 6 constructed response items scored on a four point scale. The total score gives equal nominal weight to multiple-choice and constructed-response items.
3. A test consists of 30 undifferentiated multiple-choice items and two essays, each of which is rated on a five point scale by two raters.
4. Same as 3 but, in addition, the essay total score is scaled (e.g., linearly equated) to the multiple-choice total score.

5. Same as 4, but the 30 multiple-choice items are from two categories in a table of specifications with 12 and 18 items each.
6. Same as 5 but multiple-choice scores and essay scores are equally weighted in arriving at a total raw score.
7. Same as 6 but there is a non-linear transformation of total raw scores to scale scores.

Breyer and Lewis (1994) and Livingston and Lewis (1995) provide methods that might be used to estimate single-administration decision consistency for at least some complex assessments. The Livingston and Lewis (1995) procedure is both novel and complex. It involves estimating the true score distribution using the four-parameter beta. Then the conditional distribution of scores on an alternate form (given true score) is estimated using a binomial distribution based on a novel definition of “effective test length.” For the Breyer and Lewis (1994) procedure: (a) the assessment is split into similar halves with a cut score specified for each half; (b) consistency is determined for these halves; and (c) the result is “stepped-up” to what might be expected for the full-length assessment. The Breyer and Lewis (1994) and Livingston and Lewis (1995) procedures are rather complicated, they have not been widely studied, some aspects of both methods seem rather ad hoc, and computer programs for implementing the methods are not readily available (particularly the Livingston and Lewis procedure). Of course, this does not mean the methods are necessarily flawed, but considering alternative methods seems reasonable.

At its most fundamental level, decision consistency refers to the consistency of decisions over administrations or replications of a full-length measurement procedure. It follows that there is one common feature of all methods for estimating decision consistency based on a *single* administration of an assessment—namely, implicitly or explicitly, all methods must eventually arrive at an estimate based on *hypothetical* replications of the full-length measurement procedure. The procedures that are the focus of this paper involve the following steps:

1. Generate a hypothetical full-length replication using a bootstrap procedure.
2. Apply the scoring/scaling procedure used with the original data to the bootstrap replication.
3. For each examinee, if the examinee passes for both the original data and the bootstrap replication, or fails for both the original data and the bootstrap replication, call this a conditional consistent decision (i.e., conditional on the examinee).
4. Repeat the previous step for all bootstrap replications.
5. For each examinee compute the proportion of consistent decisions over all replications.

6. Compute the average over examinees of the proportions of consistent decisions.

There are variations on these steps (particularly the last two), but these steps constitute the essence of the approach suggested here.

The first step, choosing a bootstrap sample, requires explanation, justification, and critical consideration. For example, it will be shown that, for at least some of the less complex assessments outlined previously, the use of the bootstrap in the context of the above steps gives decision-consistency results that could be obtained using the binomial and/or multinomial distributions. In such cases, the bootstrap is perhaps simpler to employ, but there is no other substantive basis for preferring it. For other, more complex assessments, the bootstrap seems considerably simpler to employ, or the binomial and multinomial distributions may not provide a sufficient basis for obtaining an estimate of decision consistency.

We begin with a brief review of the bootstrap, as it is usually discussed in statistics. Then we consider definitions of one-person-at-a-time decision consistency indices (what we call “conditional” decision-consistency indices) as well as averages of these indices over persons. Estimation issues are considered next for dichotomous data using the boot-*i* procedure and the binominal distribution, as well as polytomous data using the boot-*i* procedure and the multinomial distribution. These discussions provide the foundation for illustrating how to use the boot-*i* procedure with complex assessments such as those mentioned earlier. The Appendix provides a number of results relating to characteristics of boot-*i* replications. Many of the results reported in the Appendix inform some important comments contained in the final section of the paper.

## 1 An Overview of the Bootstrap

The bootstrap is a resampling procedure that was originally developed by Efron (1982) to assess the accuracy of a particular estimate of a parameter  $\theta$  (see, also, Efron & Tibshirani, 1986). For a statistic  $\hat{\theta}$  based on  $s$  observations, the bootstrap algorithm is based on multiple bootstrap samples, with each such sample consisting of a random sample of size  $s$  *with replacement* from the original sample. Using the bootstrap, estimation of the standard error of a statistic  $\hat{\theta}$  involves these steps:

1. using a random number generator, independently draw a large number of bootstrap samples, say  $B$  of them;
2. for each sample, evaluate the statistic of interest, say  $\hat{\theta}_b$  ( $b = 1, \dots, B$ );
3. compute the mean of the  $\hat{\theta}_b$ :

$$\hat{\theta}_B = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_b, \quad (1)$$

which is the bootstrap estimate of  $\theta$ ; and

4. compute the sample standard deviation of the  $\hat{\theta}_b$ :

$$\hat{\sigma}(\hat{\theta}_b) = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}_b - \hat{\theta}_B)^2}, \quad (2)$$

which is the bootstrap estimated standard error.

Although the bootstrap is conceptually simple, it is not unambiguously clear how to extend it to the crossed designs that have two or more dimensions or facets, such as the  $p \times i$  design, in which each person  $p$  responds to a set of items  $i$ . Brennan, Harris, and Hanson (1987) considered several possibilities, including sampling items only (boot- $i$ ), sampling persons only (boot- $p$ ), and sampling both person and items (boot- $p, i$ ) (see, also, Brennan, 2001, chap. 6). They demonstrated that each of these bootstrap procedures gives biased estimates of the variance components, as well as biased estimates of the standard errors of the estimated variance components. Subsequently Wiley (2000) derived adjustments of the various bootstrap estimators of the variance components such that the adjusted estimators are unbiased. These adjustments for the boot- $i$  procedure are discussed in Appendix A.

In this paper, the boot- $i$  procedure is used primarily as a mechanism to generate replications of the measurement procedure in order to determine the probability of a consistent decision for an examinee and for a group of examinees.<sup>1</sup> Occasional reference is made to bootstrap standard errors, but they are not the central focus here.

## 2 Decision Consistency and Boot- $i$

Throughout this section, unless otherwise noted, our focus is on a single examinee,  $p$ . For that reason, usually there is no need to use a  $p$  subscript in the notation. We generally use  $Y$  to be a random variable designating an examinee's total score over  $k$  items in a test, with  $y$  being a realization. It follows that  $\bar{y} = y/k$  is the examinee's mean score over  $k$  items.<sup>2</sup> Similarly,  $\pi$  designates the examinee's true score in the sense of the expected value of  $\bar{y}$  as  $k \rightarrow \infty$ . Given these notational conventions, a  $p$  subscript is used only when it seems helpful to improve clarity. Note that in this section, no assumptions are made about the distribution of true scores.

### 2.1 Indices

Suppose  $\lambda$  is a cut score in the total-score metric such that an examinee passes if  $y \geq \lambda$ . For an examinee, and any two randomly selected replications of the

<sup>1</sup>For estimating decision consistency only the boot- $i$  procedure seems useful.

<sup>2</sup>Note in particular that  $\bar{y}$  does *not* refer to a mean score over persons.

measurement procedure, a consistent decision is made if the examinee passes on both replications or fails on both replications. The probability of a consistent decision for an examinee is

$$\varphi_p = [\Pr(Y \geq \lambda|\pi)]^2 + [1 - \Pr(Y \geq \lambda|\pi)]^2, \quad (3)$$

which we call an index of conditional decision consistency. The probability of an inconsistent decision for an examinee is

$$1 - \varphi_p = 2 \Pr(Y \geq \lambda|\pi)[1 - \Pr(Y \geq \lambda|\pi)], \quad (4)$$

which we call an index of conditional decision *in*consistency. We call these indices “conditional” in the sense that they are conditional on the person, which is analogous to the use of the phrase “conditional” standard error of measurement.

Strictly speaking, Equations 3 and 4 for  $\varphi_p$  and  $1 - \varphi_p$ , respectively, depend on  $\pi$ , which is never known. Often,  $\bar{y}$  is used as an estimate of  $\pi$  to obtain  $\hat{\varphi}_p$  and  $1 - \hat{\varphi}_p$ . The status of these estimates of  $\varphi_p$  and  $1 - \varphi_p$  is potentially ambiguous, however.<sup>3</sup> Clearly, once  $y$  is determined for a particular test, it is known whether the examinee either passes *or* fails on that test. For an examinee who gets a passing observed score on a particular test, the only consistent decision that can be made is a passing decision on a replication; for an examinee who gets a failing observed score on a particular test, the only consistent decision that can be made is a failing decision on a replication. From this perspective, it is sometimes sensible to consider the probability of a consistent decision as

$$\varphi'_p = \begin{cases} \Pr(Y \geq \lambda|\bar{y} = \hat{\pi}) & \text{if examinee passes on actual test} \\ 1 - \Pr(Y \geq \lambda|\bar{y} = \hat{\pi}) & \text{if examinee fails on actual test.} \end{cases} \quad (5)$$

Similarly, from this perspective, the estimated probability of an inconsistent decision is

$$1 - \varphi'_p = \begin{cases} 1 - \Pr(Y \geq \lambda|\bar{y} = \hat{\pi}) & \text{if examinee passes on actual test} \\ \Pr(Y \geq \lambda|\bar{y} = \hat{\pi}) & \text{if examinee fails on actual test.} \end{cases} \quad (6)$$

As an overall index of decision consistency for a group of  $n$  examinees, we can use

$$\varphi = \frac{1}{n} \sum_{p=1}^n \varphi_p; \quad (7)$$

the corresponding overall index of decision inconsistency is  $1 - \varphi$ . Another possible index of decision consistency for a group of  $n$  examinees is

$$\varphi' = \frac{1}{n} \sum_{p=1}^n \varphi'_p; \quad (8)$$

the corresponding overall index of decision inconsistency is  $1 - \varphi'$ . As will be evident soon, using boot- $i$  replications, it is more natural (and perhaps more meaningful) to use the  $\varphi'_p$  and  $\varphi'$  statistics.

<sup>3</sup>Obviously, if  $\pi$  were known, the examinee’s pass/fail status would be known with certainty, and there would be no reason to examine issues of decision consistency.

## 2.2 Boot- $i$ Steps

Let an undifferentiated set of items be a set of items that are treated interchangeably. Usually, at a minimum, this means that the items all have the same number of possible scores (e.g., dichotomously-scored items, or polytomously-scored items with the same number of score points), but that need not be the case. For an undifferentiated set of items, the following steps can be used to estimate  $\varphi'_p$ :

1. Determine if the examinee passes or fails for the original data.
2. Get a boot- $i$  sample.
3. Determine if the examinee passes or fails for the boot- $i$  data.
4. If the examinee passes for both the original data and the boot- $i$  sample, or fails for both the original data and the boot- $i$  sample, call this a consistent decision, and code it as  $d_b = 1$ . For an inconsistent decision  $d_b = 0$ .
5. Do steps 2 and 3 a large number of times, say  $B$ , where  $b = 1, 2, \dots, B$ .
6. Then

$$\hat{\varphi}'_p = \sum_{b=1}^B d_b / B. \quad (9)$$

These steps have been specified for a specific person. There are two ways we might conceptualize repeating these steps for each of the  $n$  persons in a dataset:

- use different boot- $i$  samples for all persons or
- use the same boot- $i$  samples for each person.

Strictly speaking, using “different boot- $i$  samples” acts as if each person takes a different test for each replication (the  $i:p$  design in the terminology of generalizability theory), whereas using “the same boot- $i$  samples” acts as if each person takes the same test for each replication (the  $p \times i$  design in the terminology of generalizability theory). However, as  $B \rightarrow \infty$ , the two procedures give the same results.<sup>4</sup> Unless otherwise stated, in this paper we will generally use the “same boot- $i$  samples” conceptualization.

Given  $\hat{\varphi}'_p$ , the proportion of consistent decisions in Equation 3 can be estimated as:

$$\hat{\varphi}_p = (\hat{\varphi}'_p)^2 + (1 - \hat{\varphi}'_p)^2. \quad (10)$$

In effect,  $\hat{\varphi}_p$  “counts” both twice-passing *and* twice-failing decisions as consistent decisions, whereas  $\hat{\varphi}'_p$  “counts” only twice-passing *or* twice-failing decisions depending on the examinee’s status on the original test. The two statistics,  $\hat{\varphi}_p$  and  $\hat{\varphi}'_p$ , answer different questions about consistency. We sometimes call  $\hat{\varphi}_p$  a measure of “bi-consistency.” Note that,

<sup>4</sup>This is directly analogous to the fact that  $\sigma^2(\Delta)$  in generalizability theory is identical for the  $I:p$  and  $p \times I$  designs.

- if  $\hat{\varphi}'_p < .5$  then  $\hat{\varphi}'_p < \hat{\varphi}_p$ , and
- if  $\hat{\varphi}'_p > .5$  then  $\hat{\varphi}'_p > \hat{\varphi}_p$ .

As discussed next, as  $B \rightarrow \infty$ , these basic steps for obtaining an estimate of  $\varphi'_p$  for a specific person give the same results as those obtained using the binomial model when all the items are scored dichotomously, or the multinomial model when all the items are polytomous with the same set of score categories.

### 2.3 Dichotomous Data and the Binomial Model

Under a random sampling model with dichotomous data, the probability of obtaining any particular total score is given by the binomial probability density function:

$$\Pr(Y = y|k, \pi) = \frac{k!}{y!(k-y)!} \pi^y (1-\pi)^{k-y}. \quad (11)$$

It follows that

$$\Pr(Y \geq \lambda|k, \pi) = \sum_{y=\lambda}^k \frac{k!}{y!(k-y)!} \pi^y (1-\pi)^{k-y}, \quad (12)$$

which can be used to obtain  $\varphi_p$ , given by Equation 3.

Obviously, Equation 12 cannot be used directly because we do not know  $\pi$ . However, when  $\pi$  is replaced by the unbiased estimate  $\bar{y}$ , and Equation 12 is then used in Equation 5, the resulting estimate of  $\varphi'_p$  is identical to that obtained using the boot- $i$  procedure as  $B \rightarrow \infty$ . This is immediately obvious from the fact that boot- $i$  sampling with dichotomous data is effectively the same sampling plan that gives rise to the binomial distribution when  $\bar{y}$  is used in place of  $\pi$ .

### 2.4 Polytomous Data and the Multinomial Model

Suppose a test (or test section) consists of  $k$  polytomous items, and each item is scored as one of  $h$  possible score points,  $c_1 < c_2 < \dots < c_h$ . Assume a sample of  $k$  items is drawn at random from a universe of items, and let  $\boldsymbol{\pi} = \{\pi_1, \pi_2, \dots, \pi_h\}$  denote the proportions of items in the universe for which an examinee would get scores of  $c_1, c_2, \dots, c_h$ , respectively. Further, let  $Y_1, Y_2, \dots, Y_h$  be random variables representing the number of items scored  $c_1, c_2, \dots, c_h$ , respectively, such that  $Y_1 + Y_2 + \dots + Y_h = k$ . It follows that  $Y = c_1 Y_1 + c_2 Y_2 + \dots + c_h Y_h$  is the total raw score. Note that  $Y_1, Y_2, \dots, Y_h$  are random variables that have a multinomial distribution:

$$\Pr(Y_1 = y_1, Y_2 = y_2, \dots, Y_h = y_h, |k, \boldsymbol{\pi}) = \frac{k!}{y_1! y_2! \dots y_h!} \pi_1^{y_1} \pi_2^{y_2} \dots \pi_h^{y_h}. \quad (13)$$

This description of the multinomial model mirrors that provided by Lee (2001), who uses this model in an extensive discussion of conditional SEMs for both raw scores and scale scores.

Note that Equation 13 is for a single examinee with  $\boldsymbol{\pi}$  being that examinee's vector of true-scores (in the mean-score metric) for each of the  $h$  categories. It follows that, when  $h = 2$ , the multinomial model is identical to the binomial model for a single examinee, that was discussed in the previous section. For the binomial model, the score categories are simply  $c_1 = 0$  and  $c_2 = 1$ .

There are a number of sets of values for  $y_1, y_2, \dots, y_h$  that give a particular  $Y = y$ . So, in general, using Equation 13, the probability of a particular  $Y = y$  score is:

$$\Pr(Y = y|k, \boldsymbol{\pi}) = \sum_{c_1 y_1 + c_2 y_2 + \dots + c_h y_h = y} \Pr(Y_1 = y_1, Y_2 = y_2, \dots, Y_h = y_h, |k, \boldsymbol{\pi}), \quad (14)$$

where the sum is taken over all values of  $y_1, y_2, \dots, y_h$  that lead to  $y$ . It follows that the probability of passing is:

$$\Pr(Y \geq \lambda|k, \boldsymbol{\pi}) = \sum_{y=\lambda}^k \Pr(Y = y|k, \boldsymbol{\pi}), \quad (15)$$

where  $\Pr(Y = y|k, \boldsymbol{\pi})$  is given by Equation 14.

If  $\boldsymbol{\pi}$  were known, then Equations 3 and 7 could be used directly to obtain the conditional decision-consistency indexes  $\varphi_p$  and the overall decision-consistency index  $\varphi$ , respectively. Similarly, Equations 5 and 8 could be used directly to obtain  $\varphi'_p$  and  $\varphi'$ , respectively. When  $\boldsymbol{\pi}$  is not known, it is natural to consider using  $\hat{\pi}_1 = \bar{y}_1, \hat{\pi}_2 = \bar{y}_2, \dots, \hat{\pi}_h = \bar{y}_h$ , where  $\bar{y}_l$  ( $l = 1, 2, \dots, h$ ) is the proportion of times (over the  $k$  items) that the person's score was  $c_l$ . The  $\bar{y}_l$  are unbiased estimators of the  $\pi_l$ , and the multinomial distribution generated using them is also generated using the boot- $i$  random sampling procedure that is discussed next.

The original data effectively define an urn-sampling model in which there are  $k$  balls, and the proportions of  $c_l$  balls (i.e., scores) is  $\bar{y}_l$  ( $l = 1, 2, \dots, h$ ). When sampling is with replacement, it is well-known that this urn-sampling model gives rise to the multinomial distribution. Since any given boot- $i$  sample is a random sample with replacement from the original data, the boot- $i$  replications also give rise to the multinomial distribution in which the  $\bar{y}_l$  play the role of the  $\pi_l$ . That is, any given boot- $i$  replication is equivalent to  $k$  draws from the "multinomial urn." An infinite number of such sets of draws, grouped according to the realizations of  $Y_1, Y_2, \dots, Y_h$ , gives rise to the multinomial distribution.

## 2.5 Parametric Boot- $i$ Procedure

We can also conceptualize the boot- $i$  procedure from a parametric perspective. For example, assuming the binomial model holds, and using the original data to estimate  $\pi_l$ , a boot- $i$  sample can be generated in the following manner:

- Draw a uniform random number, say  $u$ .
- If  $u \leq \hat{\pi}$ , set the item response to 1.

- If  $u > \hat{\pi}$ , set the item response to 0.
- Repeat these steps  $k$  times.

If  $\bar{y} = \hat{\pi}$  and  $B \rightarrow \infty$ , then using this method of generating boot- $i$  replications will give the same result for  $\hat{\varphi}'_p$  as sampling with replacement from the original data, which in turn will give the same result as using Equation 12 directly.

The parametric bootstrap does have an advantage over the ordinary bootstrap in some cases, however. For example, for dichotomously-scored items, Subkoviak claims that better estimates of decision consistency might be obtained by using regressed score estimates of  $\pi$  rather than the persons' observed mean scores. A regressed score estimate of  $\pi$  is

$$\hat{\mu}_p = (1 - \rho_{YY'})\bar{\bar{y}} + \rho_{YY'}(\bar{y}_p), \quad (16)$$

where  $\rho_{YY'}$  is reliability, the subscript  $p$  is used for clarity, and  $\bar{\bar{y}}$  is the grand mean over persons and items.<sup>5</sup> The ordinary bootstrap cannot use  $\hat{\mu}_p$  as an estimate of  $\pi$ , but the parametric bootstrap can, as can Equation 12, of course.

If we assume the multinomial model holds, then the parametric bootstrap involves generating a boot- $i$  sample in the following manner:

- Draw a uniform random number, say  $u$ .
- If  $u$  is in the interval  $[0, \hat{\pi}_1]$ , set the item response to  $c_1$ .
- If  $u$  is in the interval  $(\hat{\pi}_1, \hat{\pi}_1 + \hat{\pi}_2]$ , set the item response to  $c_2$ .
- ...
- If  $u$  is in the interval  $(\sum_{l=1}^{h-1} \hat{\pi}_l, 1]$ , set the item response to  $c_h$ .
- Repeat these steps  $k$  times.

If  $\bar{y}_l = \hat{\pi}_l$  for  $l = 1, 2, \dots, h$  and  $B \rightarrow \infty$ , then using this method of generating boot- $i$  replications will give the same result for  $\hat{\varphi}'_p$  as sampling with replacement from the original data, which in turn will give the same result as using Equation 15 directly.

There is no multinomial-model regressed-score estimation procedure known to this author. That is, there is no literature known to this author that involves regressed-score estimates of the elements in  $\pi$ .

### 3 Complex Assessments and Stratified Boot- $i$

Section 2 describes three procedures for estimating decision consistency for an examinee:

<sup>5</sup>Regressed score estimates involve a number of considerations, some of which are contentious. One debatable issue is what estimate of reliability to use. Other issues are considered later.

1. the boot- $i$  procedure using sampling with replacement from the examinee's original vector of item responses;
2. the parametric boot- $i$  procedure in which estimates of  $\pi$  (for the binomial) or  $\boldsymbol{\pi}$  (for the multinomial) are used, along with a uniform random number generator, to create bootstrap item response vectors; and
3. direct use of distribution functions for the binomial or multinomial.

The procedures have been described for the simple case of a test that is viewed as consisting of a set of undifferentiated items. Indeed, we have made matters even simpler by assuming throughout most of Section 2 that a test consists of either dichotomous items or polytomous items with exactly  $h$  categories. For such simple tests, there is no compelling theoretical reason to use either bootstrap procedure to compute decision consistency indices; the distribution functions can be used directly. Probably the only theoretical reason for preferring one of the bootstrap procedures is that it could be used to estimate standard errors, although that is not the focus of this paper. Of course, from the point of view of computational ease, at least with a computer, one of the bootstrap procedures may be preferable.

For complex assessments such as the illustrative examples listed on pages 1–2, however, direct use of distribution functions is quite difficult and often virtually impossible. By contrast, it is relatively straightforward to apply the boot- $i$  procedure or its parametric counterpart. With complex assessments, instead of using a single boot- $i$  sample to simulate a replication, we use as many boot- $i$  samples as are built into the design of the assessment. In this sense, we replicate complex assessments through a stratified boot- $i$  sampling procedure.

### 3.1 A Synthetic Example

Consider a somewhat more detailed version of the second example on page 1:

A test consists of 40 undifferentiated multiple-choice items, four constructed response items scored on a three point scale (0, 1, 2), and six constructed response items scored on a four point scale (0, 1, 2, 3). The total score, expressed as a percent, gives equal nominal weight to multiple-choice and constructed-response items, with the cut score being 65%.

Suppose an examinee's response vector for the original data is:

Item	11111111112222222222333333333334	4444	444445
Numbers	1234567890123456789012345678901234567890	1234	567890
Responses	1011001001001110111100010101011000100110	1222	012323

For this examinee, the three raw scores are 20, 7, and 11, respectively. Therefore, the multiple-choice raw score is 20, and the constructed response raw score is

18. It follows that the multiple-choice percent is  $100(20/40) = 50.0\%$ , the constructed response percent is  $100(18/26) = 69.2\%$ , weighing them equally gives a score of 59.6%, and the examinee fails.

Now, suppose we take a stratified random sample with replacement from the examinee's response vector. That is, suppose we take a random sample with replacement from the examinee's vector of responses to the 40 multiple-choice items. Then, we take a random sample with replacement from the examinee's responses to the four three-point constructed response items. Finally, we take a random sample with replacement from the examinee's responses to the six four-point constructed response items. Suppose the examinee's resulting stratified boot- $i$  response vector is:

Item	11231	33333133	33123112	212233	212211	4444	444445
Numbers	2337024610206813955667486742469128701811	4214	677980				
Responses	0100101011010100101000010010110100010100	2212	122233				

For this boot- $i$  sample, the three raw scores are 16, 7, and 13, respectively. Therefore, the multiple-choice raw score is 16, and the constructed response raw score is 20. It follows that the multiple-choice percent is  $100(16/40) = 40.0\%$ , the constructed response percent is  $100(20/26) = 76.9\%$ , weighing them equally gives a score of 58.5%, and the examinee fails. Since the examinee fails for both the original data and the boot- $i$  sample, a consistent decision is made. Repeating this process a large number of times (say  $B$ ) gives an estimated proportion of consistent decisions for this examinee,  $\hat{\phi}'_p$ .

To get an estimated proportion of consistent decisions over all tested examinees,  $\hat{\phi}'$ , we can use the same  $B$  stratified boot- $i$  sampled items with all examinees<sup>6</sup> and average the  $n$  values of  $\hat{\phi}'_p$ . To get the proportion of bi-consistent decisions,  $\hat{\phi}_p$ , we simply use Equation 10, and averaging over persons we obtain  $\hat{\phi}$ .

### 3.2 A Semi-real Example

A particular licensure/certification testing program consists of 80 multiple-choice items and eight constructed-response items that are scored on a four-point scale of 1–4. The multiple-choice raw score is converted to a scale-score,  $SS_{mc}$ , using a linear transformation; and the constructed-response raw score is also converted to a scale-score,  $SS_{cr}$ , using a linear transformation. The composite scale score is defined as

$$SS = .6(SS_{mc}) + .4(SS_{cr}).$$

Cut scores are determined for both the multiple-choice and the constructed-response items that lead to a SS cut-score of 240.

<sup>6</sup>For example, for each examinee, the first boot- $i$  sample would be the examinee's responses to items 12, 13, 23, etc.

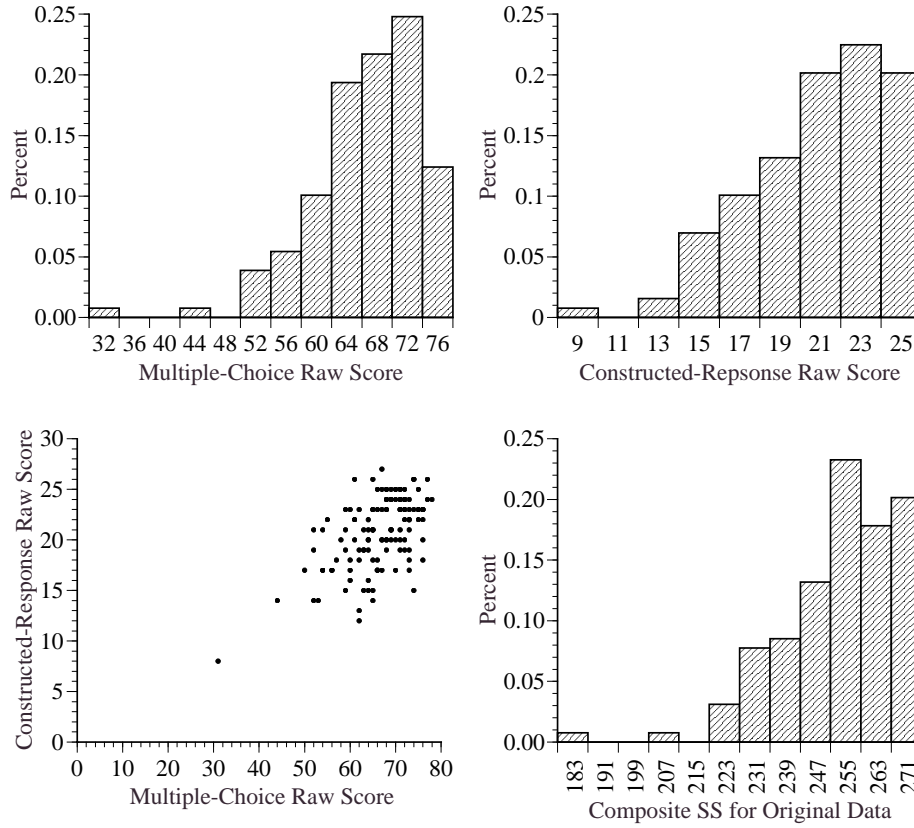


Figure 1: Multiple-choice raw scores, constructed-response raw scores, and composite standard scores for original data.

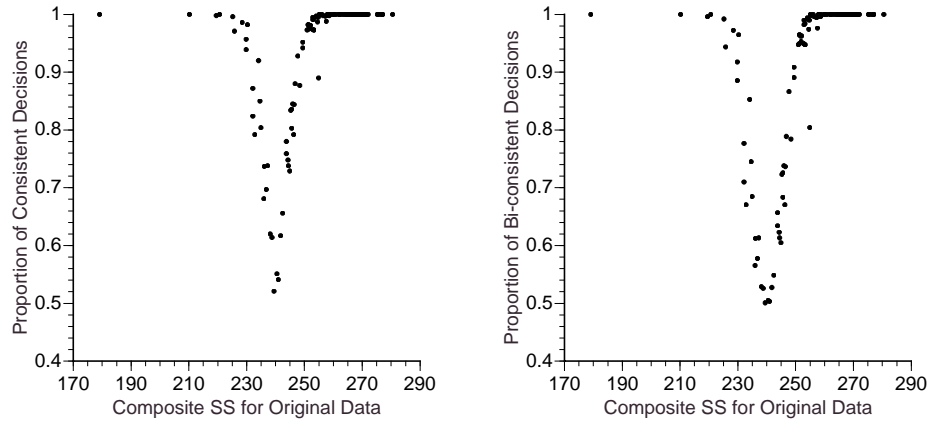


Figure 2: Proportions of consistent decisions,  $\hat{\phi}'_p$ , and bi-consistent decisions,  $\hat{\phi}_p$ .

A total of 129 examinees took a recent version of the test, which used 49 as the multiple-choice cut score and 23 as the constructed-response cut score. Figure 1 provides histograms of the the multiple-choice raw scores, the constructed response raw scores, and the composite standard scores, as well as a scatterplot of the multiple-choice and constructed response raw scores.

Using  $B = 1000$  boot- $i$  replications, the average proportion of consistent decisions is  $\hat{\varphi}' = .936$ , and the average proportion of bi-consistent decisions is  $\hat{\varphi} = .907$ . Figure 2 plots the individual  $\hat{\varphi}'_p$  and  $\hat{\varphi}_p$  values for each person. Using Equation 2 with  $\hat{\theta}_b$  replaced by the average number of consistent decisions for each replication  $b$ , and  $\hat{\theta}_B$  replaced by  $\hat{\varphi}' = .936$ , the bootstrap estimated standard error of  $\hat{\varphi}'$  is .021.<sup>7</sup>

Although the four constructed-response items are all scored using the same 1–4 scale, the first two items are of a different type from the last two. Therefore, the two types might be treated separately for boot- $i$  sampling purposes. That is, for any replication, a stratified boot- $i$  sample could consist of a random sample with replacement from the 80 multiple-choice items, a random sample with replacement from the four scores associated with the first two constructed-response items, and a random sample with replacement from the four scores associated with the last two constructed-response items. Using this definition of a stratified boot- $i$  sample, the average proportion of consistent decisions is  $\hat{\varphi}' = .940$  with a bootstrap estimated standard error of .021, and the average proportion of bi-consistent decisions is  $\hat{\varphi} = .913$ .

For many testing programs, the various forms administered on different test dates are equated to each other. Under such circumstances, it seems likely that investigators will be interested in decision consistency for hypothetical forms that are equated to the original form.<sup>8</sup> This is easily accomplished using boot- $i$  samples as forms. For the example considered here, using three-strata boot- $i$  samples and linearly equating the observed scores for each of them to the observed scores for the original data, we obtain an average proportion of consistent decisions of  $\hat{\varphi}' = .945$  with a standard error of .019, and an average proportion of bi-consistent decisions of  $\hat{\varphi} = .921$ .

The decision-consistency results for this example are summarized in the following table. For the three-strata boot- $i$  sample,  $\hat{\varphi}'$  is necessarily larger (although not by much in this case) than for the two-strata sample because the additional level of stratification increases the likelihood that the same constructed-response scores will be included in different boot- $i$  samples. To appreciate this, it is helpful to realize that if every constructed-response score were asso-

<sup>7</sup>There is no corresponding bootstrap estimated standard error of  $\hat{\varphi} = .907$ , which is the average over persons of the estimates of  $\hat{\varphi}_p$ , with each such estimate being a mean over replications. It is possible to get  $\hat{\varphi}$  for each replication, say  $\hat{\varphi}_b$ , by computing  $\hat{\varphi} = (\hat{\varphi}')^2 + (1 - \hat{\varphi}')^2$  for each replication. Then, the standard deviation of these values, over replications, is an estimated bootstrap standard error. For these data, the mean over replications of  $\hat{\varphi}_b$ , say  $\hat{\varphi}_B$ , is .881, and the estimated bootstrap standard error is .037. Note that  $\hat{\varphi}_B$  is different from  $\hat{\varphi} = .907$ , discussed above.

<sup>8</sup>In fact, ideally we would like to administer two forms to all examinees and estimate decision consistency directly. If this were possible, the methodology in this paper would be unnecessary.

Conditions	$\hat{\varphi}'$	$\hat{\varphi}$
Two-strata boot- <i>i</i>	.936	.907
Three-strata boot- <i>i</i>	.940	.913
Three-strata boot- <i>i</i> with equating	.945	.921

ciated with its own stratum, then every boot-*i* sample would include the same constructed-response scores. Equating tends to increase decision consistency because equated scores are, on average, more similar to scores for the original data.

For these data, using a particular half-test split, the Breyer and Lewis (1994) procedure gave an estimated decision-consistency index of .865. Both conceptually and numerically, the closest boot-*i* result is  $\hat{\varphi} = .907$  using two strata without equating. Still, the two values are noticeably different. There are probably a number of reasons for the difference. For example, the Breyer and Lewis procedure is dependent on how the assessment is split into two halves. Splitting a complex assessment into “equivalent” halves is likely to be only approximately attainable. Also, the Breyer and Lewis procedure depends on a set of rather stringent assumptions for “stepping up” the decision consistency for the half tests to that for the full-length test. On the other hand, the boot-*i* procedure makes its own assumptions that may not be fully acceptable. For example, in the next section it is suggested that the boot-*i* procedure might give a somewhat inflated estimate of decision consistency.

## 4 Discussion and Conclusions

For an assessment that consists of an undifferentiated set of items with the same number of score categories (e.g., dichotomously-scored items), relatively straightforward procedures exist for estimating decision consistency based on a single administration of the assessment. For complex assessments, however, few procedures exist. The boot-*i* procedure proposed in this paper is relatively simple to implement, and it can be used no matter how complicated the assessment may be.

Many procedures for estimating decision consistency require assumptions about the underlying distribution of true scores. For example, beta-binomial procedures assume that true scores have a two- or four-parameter beta distribution. The fact that the boot-*i* procedure requires no assumptions about the distribution of true scores makes the procedure generally applicable and computationally straightforward. However, investigators pay a price when true scores are ignored. Clearly, if the distribution of true scores were known, or at least well-estimated, then we could estimate decision consistency for examinees conditional on true score, rather than using observed scores as estimates of true scores. However, the absence of true scores is not a fatal problem as long as we

are willing to accept the average of individual estimates of decision consistency as an overall estimate of decision consistency. This is much like accepting the average of conditional error variances as the overall error variance, which is done routinely in both classical theory and generalizability theory.

The consistency statistic  $\varphi'_p$  and bi-consistency statistic  $\varphi_p$  are related as follows:

$$\varphi_p = (\varphi'_p)^2 + (1 - \varphi'_p)^2.$$

When we focus on an individual examinee, once the score on the original test is determined, the examinee either passes or fails. For a passing examinee, the only possible consistent decision is pass-pass; for a failing examinee, the only possible consistent decision is fail-fail. In this sense, the consistency statistic  $\varphi'_p$  is sensible and the bi-consistency statistic  $\varphi_p$  is not; similarly, the overall consistency statistic  $\varphi'$  is sensible and  $\varphi$  is not. However, virtually all of the decision-consistency literature, except this paper, uses  $\varphi_p$  and/or  $\varphi$ . Why the difference? The essential reason is that most procedures do not take the original, observed data as definitive of pass/fail status. Rather, through the assumption of a true score distribution or splitting a test into halves, most procedures generate a bivariate distribution of *expected* observed scores from which  $\varphi$  is estimated directly.

In the context of this paper, is there a sense in which the bi-consistency statistics  $\varphi_p$  and  $\varphi$  can be defended? Yes, but in a somewhat loose sense. Consider the simple case of dichotomous items and the binomial distribution. When we use  $\bar{y}$  as an estimate of  $\pi$  for an individual, and act as if the estimate really is  $\pi$ , then  $\hat{\varphi}_p$  (and hence  $\hat{\varphi}$ ) is defensible. This “acting as if” scenario is probably more credible if we use a regressed-score estimate of  $\pi$ , rather than  $\bar{y}$ , since the distribution of regressed-score estimates is closer to the distribution of true scores than is the distribution of observed scores. Unfortunately, however, there is no obvious way to use regressed-score estimates in the boot-*i* procedure with complex assessments.

The appendix to this paper provides a number of results about boot-*i* replications, where a particular boot-*i* replication is to be viewed as using the same (non-stratified) boot-*i* sample of items for each person. It is evident from these results that, when  $B$  is finite, the original data and a boot-*i* sample are more similar, in some respects, than are randomly parallel forms. Therefore, if we take randomly parallel forms as a definition of “equivalent” forms, it may be that boot-*i* measures of decision consistency will be inflated, at least somewhat, relative to what such measures would be for randomly parallel forms. From this perspective, one might argue that there isn’t enough noise in boot-*i* replications. However, there is a kind of precedent for using the boot-*i* procedure in measurement contexts. As discussed in Section A.6, for an undifferentiated set of  $k$  items, the estimated conditional absolute SEM in generalizability theory,  $\hat{\sigma}(\Delta_p)$ , is directly related to the boot-*i* standard error of the mean for a person, namely,

$$\hat{\sigma}(\Delta_p) = \sqrt{\frac{k}{k-1}} \hat{\sigma}_b(\bar{Y}_p).$$

On balance, boot-*i* estimates of decision consistency seem useful if they are interpreted with appropriate cautions, and they seem particularly valuable for complex assessments, especially when equating is employed in an operational testing program.

## 5 References

- Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag.
- Brennan, R. L., Harris, D. J., & Hanson, B. A. (1987). *The bootstrap and other procedures for examining the variability of estimated variance components in testing contexts* (American College Testing Research Report No. 87-7). Iowa City, IA: ACT, Inc.
- Breyer, F. J., & Lewis, C. (1994). *Pass-fail reliability for tests with cut scores: A simplified method* (ETS Research Report No. 94-39) Princeton, NJ: Educational Testing Service.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
- Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Philadelphia: SIAM.
- Efron, B. & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1, 54-77.
- Feldt, L. S. & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 105-146). New York: American Council on Education and MacMillan.
- Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement*, 27, 345-359.
- Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, 13(4), 253-264.
- Lee, W. (April, 2001). *A multinomial error model for tests with polytomous items*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32(2), 179-197.

- Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26*(4), 412–432.
- Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13*, 265–276.
- Wiley, E. W. (2000). *Bootstrap strategies for variance component estimation: Theoretical and empirical results*. Unpublished doctoral dissertation, Stanford.

## A Some Characteristics of boot- $i$ Replications

In the body of this paper, the primary focus is on a single person, which permits a simplification of notational conventions. Here, except in Section A.6, we focus on the person-by-item matrix that results from using the same boot- $i$  sample of items for all persons. Furthermore, we need to carefully distinguish between the total-score metric and the mean score metric. Hence, the notation is more complicated and (admittedly) less elegant.<sup>9</sup> Let

- $n$  = number of persons,
- $k$  = number of items,
- $Y_{pi}$  = score for person  $p$  on item  $i$  for the original data,
- $X_{pj}$  =  $j$ -th score for person  $p$  based on a particular boot- $i$  sample,
- $Y_p$  = total score for person  $p$  for the original data,
- $X_p$  = total score for person  $p$  for a particular boot- $i$  sample,
- $\bar{Y}_p$  = mean score (over  $k$  items) for person  $p$  for the original data,
- $\bar{X}_p$  = mean score (over  $k$  items) for person  $p$  for a particular boot- $i$  sample,
- $\sigma^2(\alpha)$  = the random effects variance component for the effect  $\alpha$  for the original data,
- $\sigma^2(\alpha|b)$  = the random effects variance component for the effect  $\alpha$  for a boot- $i$  sample,
- $EMS(\alpha)$  = the expected mean square under the random model for the effect  $\alpha$  for the original data, and
- $EMS(\alpha|b)$  = the expected mean square under the random model for the effect  $\alpha$  for a boot- $i$  sample.

In the following sections, the boot- $i$  sampling procedure is sometimes simply designated  $b$ .

---

<sup>9</sup>For example, in this appendix usually the notation does not distinguish between a random variable and its realization.

## A.1 EMS Equations and Estimated Variance Components

The random-model expected-mean-square equations for the original data  $Y$  are:

$$\mathbf{EMS}(p) = \sigma^2(pi) + k\sigma^2(p) \quad (17)$$

$$\mathbf{EMS}(i) = \sigma^2(pi) + n\sigma^2(i) \quad (18)$$

$$\mathbf{EMS}(pi) = \sigma^2(pi). \quad (19)$$

Using the  $\mathbf{EMS}$  equations, the well-known estimators of the variance components are:

$$\hat{\sigma}^2(p) = \frac{MS(p) - MS(pi)}{k} \quad (20)$$

$$\hat{\sigma}^2(i) = \frac{MS(i) - MS(pi)}{n} \quad (21)$$

$$\hat{\sigma}^2(pi) = MS(pi). \quad (22)$$

Similarly, under the boot- $i$  procedure, the expected-mean-square equations for the boot- $i$   $X$  data are:

$$\mathbf{EMS}(p|b) = \sigma^2(pi|b) + k\sigma^2(p|b) \quad (23)$$

$$\mathbf{EMS}(i|b) = \sigma^2(pi|b) + n\sigma^2(i|b) \quad (24)$$

$$\mathbf{EMS}(pi|b) = \sigma^2(pi|b). \quad (25)$$

Using the  $\mathbf{EMS}$  equations, the unadjusted bootstrap estimators of the variance components are:

$$\hat{\sigma}^2(p|b) = \frac{MS(p|b) - MS(pi|b)}{k} \quad (26)$$

$$\hat{\sigma}^2(i|b) = \frac{MS(i|b) - MS(pi|b)}{n} \quad (27)$$

$$\hat{\sigma}^2(pi|b) = MS(pi|b). \quad (28)$$

## A.2 Bias-Correcting Adjustments

Wiley (2000) derived the following bias-correcting adjustments for the  $\hat{\sigma}^2(\alpha|b)$ :

$$\hat{\sigma}^2(p) = \hat{\sigma}^2(p|b) - \left(\frac{1}{k-1}\right) \hat{\sigma}^2(pi|b) \quad (29)$$

$$\hat{\sigma}^2(i) = \left(\frac{k}{k-1}\right) \hat{\sigma}^2(i|b) \quad (30)$$

$$\hat{\sigma}^2(pi) = \left(\frac{k}{k-1}\right) \hat{\sigma}^2(pi|b). \quad (31)$$

It is easy to demonstrate that these adjustments imply that:

$$\mathbf{EMS}(p|b) = \mathbf{EMS}(p) + \left(\frac{k-1}{k}\right) \mathbf{EMS}(pi) \quad (32)$$

$$\mathbf{E}MS(i|b) = \left(\frac{k-1}{k}\right) \mathbf{E}MS(i) \quad (33)$$

$$\mathbf{E}MS(pi|b) = \left(\frac{k-1}{k}\right) \mathbf{E}MS(pi). \quad (34)$$

### A.3 Expected Variance of Boot- $i$ Person Scores

Note that  $\sigma^2(\bar{X}_p) = MS(p|b)/k$ , where  $\sigma^2(\bar{X}_p)$  is computed using a divisor of  $n - 1$ . Replacing Equations 17 and 19 in the expression of  $\mathbf{E}MS(p|b)$  in Equation 32 gives the expected value of  $\sigma^2(\bar{X}_p)$  over boot- $i$  replications.

$$\begin{aligned} \mathbf{E} \sigma^2(\bar{X}_p) &= \frac{\mathbf{E}MS(p|b)}{k} \\ &= \frac{\sigma^2(pi) + k \sigma^2(p)}{k} + \left(\frac{k-1}{k}\right) \frac{\sigma^2(pi)}{k} \\ &= \sigma^2(\bar{Y}_p) + \left(\frac{k-1}{k}\right) \frac{\sigma^2(pi)}{k}. \end{aligned} \quad (35)$$

In words, the expected value of the variance of the observed mean scores for a boot- $i$  sample is larger than the expected value of the variance of the observed mean scores for the original data by  $(k-1)/k$  times relative error variance for the original data. Letting  $Y_p = k\bar{Y}_p$  and  $X_p = k\bar{X}_p$  designate person total scores for the original and boot- $i$  data, respectively, the total-score version of Equation 35 is obtained by multiplying both sides by  $k^2$ , which gives

$$\mathbf{E} \sigma^2(X_p) = \sigma^2(Y_p) + (k-1)\sigma^2(pi). \quad (36)$$

### A.4 Covariance of Original-Data and Boot- $i$ Person Scores

The original-data and boot- $i$  total scores can be represented as

$$Y_p = Y_{p1} + Y_{p2} + \cdots + Y_{pi} + \cdots + Y_{pk}$$

and

$$X_p = X_{p1} + X_{p2} + \cdots + X_{pj} + \cdots + X_{pk}.$$

Note that the subscript numbers  $(1, 2, \dots, k)$  designate ordinal positions (e.g.,  $Y_{p2}$  is the score for person  $p$  on the *second* item) in the item-score vectors. In addition, for  $Y$  if  $i \neq i'$  then the items are different. By contrast, for  $X$  if  $j \neq j'$ , the items can be the same. To draw attention to this distinction,  $i$  and  $i'$  will be used exclusively with  $Y$ , and  $j$  and  $j'$  will be used exclusively with  $X$ . We wish to determine the expected value of the covariance between  $Y_p$  and  $X_p$ :

$$\mathbf{E} \sigma(Y_p, X_p) = \sum_i \sum_j \mathbf{E} \sigma(Y_i, X_j) = \sum_i \mathbf{E} \sigma(Y_i, \sum_j X_j), \quad (37)$$

where the expected value is taken over boot- $i$  samples, and we have suppressed the person subscript in  $Y_i$  and  $X_j$  to simplify notation.

For any particular  $i$  and  $j$  (pair of ordinal positions),

$$\mathbf{E} \sigma(Y_i, X_j) = \frac{1}{n} \sum_{i'=1}^k \sigma(Y_i, Y_{i'}). \quad (38)$$

To understand this equation, recall the  $Y_i$  is the score for the specific item in the  $i$ -th position; by contrast, the actual item that gives rise to  $X_j$  varies over boot- $i$  samples, with each item occurring in the  $j$ -th position an equal number of times. Consider an example. Suppose, that  $Y$  is the sum of scores for three items,  $a$ ,  $b$ , and  $c$ . Clearly, there are  $3 \times 3 \times 3 = 27$  possible ordered versions of  $X$ :

$$\begin{array}{ccccccccc} a a a & a a b & a a c & a b a & a b b & a b c & a c a & a c b & a c c \\ b a a & b a b & b a c & b b a & b b b & b b c & b c a & b c b & b c c \\ c a a & c a b & c a c & c b a & c b b & c b c & c c a & c c b & c c c \end{array}$$

Since each of these possibilities for  $X$  is equally likely, the probability is exactly  $1/3$  that each individual item will occur in the first position, the second position, or the third position.

Now, note that the righthand side of Equation 38 is constant for all values of  $j$ . It follows that the expected covariance between  $Y_i$  and the boot- $i$  total score is

$$\mathbf{E} \sigma(Y_i, \sum_j X_j) = k \left[ \frac{1}{k} \sum_{i'=1}^k \sigma(Y_i, Y_{i'}) \right] = \sum_{i'=1}^k \sigma(Y_i, Y_{i'}).$$

Replacing this result in Equation 37 gives

$$\begin{aligned} \mathbf{E} \sigma(Y_p, X_p) &= \sum_i \mathbf{E} \sigma(Y_i, \sum_j X_j) \\ &= \sum_i \sum_{i'} \sigma(Y_i, Y_{i'}) \\ &= \sigma^2(Y_p). \end{aligned} \quad (39)$$

In words, the expected covariance between total scores for the original data and boot- $i$  data is the variance of the total scores for the original data. Similarly, in the mean-score metric,

$$\mathbf{E} \sigma(\bar{Y}_p, \bar{X}_p) = \sigma^2(\bar{Y}_p). \quad (40)$$

## A.5 Parallel Forms and the Boot- $i$ Procedure

For classically parallel forms, it is assumed that the observed-score means, variances, and covariances are equal. In addition, it can be shown that the covariance between classically parallel forms equals true score variance. For the assumption of randomly parallel forms in generalizability theory (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972), these statements do not hold for specific forms or pairs of forms, but they do hold for expected values over forms or pairs of forms. So, for example, in generalizability theory, the expected

value over pairs of forms of the observed covariance (in the mean-score metric) is universe score variance.

The expected value of the mean over boot- $i$  replications,  $X$ , equals the mean for the original data,  $Y$ . By contrast, we have shown that the expected value of the variance for boot- $i$  replications,  $X$ , is larger than the variance for the original data,  $Y$  (see Equations 35 and 36), and the expected value of the covariance between  $Y$  and  $X$  is larger than true-score (or universe-score) variance (see Equations 39 and 40). It is possible, however, to transform boot- $i$  scores so that they more nearly satisfy the tenets of classically-parallel or randomly-parallel forms.

### A.5.1 Equal Means and Variances

For *any* bootstrap sample, consider the linear transformation  $X'_p = a' + b'X_p$ , where

$$b' = \frac{\sigma(Y_p)}{\sigma(X_p)} \quad \text{and} \quad a' = \mu(Y_p) - b'\mu(X_p). \quad (41)$$

Under this transformation, it is clear that  $\mu(X'_p) = \mu(Y_p)$  and  $\sigma^2(X'_p) = \sigma^2(Y_p)$ . In effect, this transformation is a linear equating of boot- $i$  total scores to the scale of  $Y$  total scores. Similarly, for the mean score metric, under the transformation  $\bar{X}'_p = a'/k + b'\bar{X}_p$ , it follows that  $\mu(\bar{X}'_p) = \mu(\bar{Y}_p)$  and  $\sigma^2(\bar{X}'_p) = \sigma^2(\bar{Y}_p)$ .

The above results give equal means and equal variances for *every* boot- $i$  sample. A weaker result, for the total score metric, that applies to a *randomly selected* boot- $i$  sample is obtained using the transformation  $X''_p = a'' + b''X_p$ , where

$$b'' = \sqrt{\frac{\sigma^2(Y_p)}{\sigma^2(Y_p) + (k-1)\sigma^2(pi)}} \quad \text{and} \quad a'' = \mu(Y_p) - b''\mu(X_p). \quad (42)$$

Under this transformation,  $\mathbf{E} \mu(X''_p) = \mu(Y_p)$  and  $\mathbf{E} \sigma^2(X''_p) = \sigma^2(Y_p)$ . That is, for a randomly selected boot- $i$  sample, it is expected that the mean of  $X''_p$  will equal the mean of  $Y_p$ , and the variance of  $X''_p$  will equal the variance of  $Y_p$ . Corresponding results apply to the mean-score metric, under the transformation  $\bar{X}''_p = a''/k + b''\bar{X}_p$ .

### A.5.2 Covariance of Original and Boot- $i$ Scores

Also, it is a simple matter to transform boot- $i$  scores so that the expected value of the covariance between the original data and the transformed boot- $i$  scores is true score variance for  $Y$ . Specifically, in the total-score metric, consider the linear transformation  $X'''_p = a''' + b'''X_p$ , where

$$b''' = \rho_{YY'} \quad \text{and} \quad a''' = \mu(Y_p) - b''' \mu(X_p). \quad (43)$$

For this transformation, using the result in Equation 39,

$$\mathbf{E} \sigma(Y_p, X'''_p) = \rho_{YY'} \mathbf{E} \sigma(Y_p, X_p) = \rho_{YY'} \sigma^2(Y_p) = \sigma^2(T_{Y_p}). \quad (44)$$

Similarly, in the mean score metric,  $\mathbf{E} \sigma(\bar{Y}_p, \bar{X}_p''') = \sigma^2(T_{\bar{Y}_p})$ .

For a randomly selected boot- $i$  sample, the  $X_p'''$  transformation can be represented as

$$X_p''' = (1 - \rho_{YY'})\mu(Y_p) + \rho_{YY'}X_p, \quad (45)$$

which has a form reminiscent of a regressed-score estimate of true scores. There is, however, an important difference. The variance of the usual regressed-score estimates is true-score variance times reliability, whereas the variance of the estimates in Equation 45 is  $\rho_{YY'}^2 \sigma^2(X_p)$ , which does not equal true-score variance times reliability for either  $X$  or  $Y$ .

### A.5.3 Impossibility of Satisfying all Conditions

There is no single linear transformation of boot- $i$  scores, say  $X_p^*$ , such that, in general,  $\sigma^2(X_p^*) = \sigma^2(Y_p)$  and  $\mathbf{E} \sigma(Y_p, X_p^*) = \sigma^2(T_{Y_p})$ . This is evident from the fact that the slope of the transformation in Equation 41, namely  $\sigma(Y_p)/\sigma(X_p)$ , does not generally equal the slope of the transformation in Equation 44, namely  $\rho_{YY'}$ . Also, these two conditions cannot be satisfied simultaneously for expected scores over boot- $i$  samples, as is evident from the fact that  $b'''$  in Equation 42 does not equal  $\rho_{YY'}$ . Indeed, it can be shown that

$$\rho_{YY'} < \sqrt{\frac{\sigma^2(Y_p)}{\sigma^2(Y_p) + (k-1)\sigma^2(pi)}}$$

which means that satisfying the condition that  $\mathbf{E} \sigma(\bar{Y}_p, \bar{X}_p''') = \sigma^2(T_{\bar{Y}_p})$  results in less shrinkage to the mean than satisfying the condition that  $\mathbf{E} \sigma^2(X_p''') = \sigma^2(Y_p)$ .

In short, boot- $i$  replications do not satisfy all the tenets of classically-parallel or randomly-parallel forms. Although various linear transformations of boot- $i$  scores give results that are more nearly parallel (classically or randomly), no single linear transformation will lead to converted boot- $i$  scores that satisfy all the tenets of parallel forms. These facts do not necessarily mean that boot- $i$  replications are fundamentally flawed; all definitions of replications are idealizations in some sense. However, the differences between boot- $i$  replications and parallel (classically or randomly) forms should be acknowledged by investigators

## A.6 Boot- $i$ Replications and Conditional SEMs

There is a close relationship between boot- $i$  replications and conditional standard errors of measurement (SEMs) in generalizability theory, which are discussed extensively by Brennan (2001). This relationship is discussed after a brief review of conditional SEMs.

In generalizability theory, for the mean score metric, absolute error for person  $p$  is  $\Delta_p = \bar{Y}_p - \mu_p$ , where  $\mu_p$  is the person's universe score (or true score). The associated error variance is

$$\sigma^2(\Delta_p) \equiv \text{var}(\bar{Y}_p - \mu_p|p). \quad (46)$$

It is the variance of the mean over  $k$  items for person  $p$ , which is called the conditional absolute error variance. The average over persons of  $\sigma^2(\Delta_p)$  is  $\sigma^2(\Delta) = [\sigma^2(i) + \sigma^2(pi)]/k$ .

An unbiased estimator of  $\sigma^2(\Delta_p)$  is

$$\hat{\sigma}^2(\Delta_p) = \frac{\text{var}(Y_{pi}|p)}{k} = \frac{\sum_i (Y_{pi} - \bar{Y}_p)^2}{k(k-1)}. \quad (47)$$

The average, over persons, of these estimates is  $\hat{\sigma}^2(\Delta) = [\hat{\sigma}^2(i) + \hat{\sigma}^2(pi)]/k$ , which is the usual estimate of  $\sigma^2(\Delta)$ .

The square root of Equation 47 provides an estimator of the conditional absolute SEM:

$$\hat{\sigma}(\Delta_p) = \sqrt{\frac{\sum_i (Y_{pi} - \bar{Y}_p)^2}{k(k-1)}}. \quad (48)$$

If items are scored dichotomously, this estimator is Lord's (1955, 1957) conditional SEM:

$$\hat{\sigma}(\Delta_p) = \sqrt{\frac{\bar{Y}_p(1 - \bar{Y}_p)}{k-1}}. \quad (49)$$

When we focus on a single person, the boot- $i$  procedure is the simple bootstrap procedure, and the conditional absolute SEM is just the standard error of the mean. Under these circumstances, Efron (1982) showed that the bootstrap estimated standard error in Equation 2 is

$$\hat{\sigma}_b(\bar{Y}_p) = \sqrt{\left(\frac{1}{k}\right) \left[\frac{\sum_i (Y_{pi} - \bar{Y}_p)^2}{k}\right]}.$$

It follows that the estimated conditional absolute SEM is

$$\hat{\sigma}(\Delta_p) = \sqrt{\frac{k}{k-1}} \hat{\sigma}_b(\bar{Y}_p).$$

Note that the above conditional SEM results are for the mean-score metric. Multiplication by  $k$  gives the corresponding results for the total-score metric.