Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 18

Correcting for Bias in Single-Administration Decision Consistency Indexes

Robert L. Brennan
Won-Chan Lee[†]

July 2006

[†]Robert L. Brennan is E. F. Lindquist Chair in Measurement and Testing and Director, Center for Advanced Studies in Measurement and Assessment (CASMA), 210D Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: robert-brennan@uiowa.edu). Won-Chan Lee is a Research Scientist in CASMA, 210E Lindquist Center, University of Iowa, Iowa City, IA 52242 (email: won-chan-lee@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
Web: www.education.uiowa.edu/casma

Contents

Abstract **iv**

1 Introduction **1**

2 Dichotomous Data and the Binomial Error Model **2**
2.1 Simulations . . . 3
2.2 An Optimal Estimate of $\boldsymbol{\pi}$ . . . 7
2.3 Simulations Revisited . . . 8
2.4 Parametric Bootstrap . . . 9
2.5 Estimating Reliability . . . 10
2.6 Other Issues . . . 10

3 Polytomous Data and the Multinomial Error Model **10**
3.1 Estimating ${}_w\hat{\pi}_l$ . . . 11
3.2 Properties of ${}_w\hat{\pi}_l$ . . . 12
3.3 Parametric Bootstrap . . . 13
3.4 Estimating Reliability . . . 13

4 Complex Assessments and the Compound Multinomial Error Model **14**
4.1 Estimation . . . 15
4.2 Parametric Bootstrap . . . 15

5 Concluding Comments **16**

6 References **17**

Abstract

Subkoviak (1976) proposed a procedure for estimating decision consistency for a group of examinees based on a single administration of a test consisting of dichotomously-scored items. A distinguishing feature of his approach is that it begins by estimating agreement for each individual examinee. To do so, Subkoviak suggested estimating an examinee's true score using either the examinee's observed mean score or the examinee's regressed-score estimate. The principal purpose of this paper is to examine the amount of bias that results from these two suggestions as well as two other estimates that are considered. In doing so, we propose an "optimally" weighted estimate of an examinee's true score that appears to reduce bias substantially. We focus principally on the dichotomous-data case, but we also consider extensions to polytomous data and to complex assessments.

1 Introduction

Thirty years ago in back-to-back papers in the same issue of the *Journal of Educational Measurement*, Huynh (1976) and Subkoviak (1976) provided procedures for estimating decision consistency using data from only a single administration of a test consisting of k dichotomously-scored items, with a number-correct cut score of λ. Huynh's procedure assumes the data conform to the assumptions of a beta-binomial model; i.e., for every proportion-correct true score (π) the conditional distribution of observed scores (X) is binomial, the regression of true scores on observed scores is linear, and the distribution of true scores is the two-parameter beta (with parameters denoted a and b). Under these assumptions, the distribution of observed scores is negative hypergeometric. Huynh's major contribution was the provision of particularly elegant recursive algorithms for computing the negative hypergeometric distribution and its bivariate counterpart, thereby making it relatively easy to estimate agreement (P), chance agreement (P_c), and kappa (κ) for the *group* of n examinees who took the test.

By contrast, Subkoviak's procedure does not require the full set of beta-binomial assumptions although it does assume that the conditional distribution of observed scores given true score is binomial; i.e,

$$\Pr(X = x|\pi) = \left(\begin{array}{c} k \\ x \end{array} \right) \pi^x (1-\pi)^{k-x}, \tag{1}$$

which is also the distribution of errors for an examinee with a true proportion-correct score of π. In addition, Subkoviak's procedure differs from Huynh's in that Subkoviak obtains group agreement by averaging examinee-level agreement. Specifically, letting p be agreement for an examinee, then group agreement is

$$P = \overline{p}, \tag{2}$$

where $\overline{p}$ is the average value of p over the n examinees.[1]

Assuming two identically distributed and independent distributions for X,

$$p = [\Pr(X \geq \lambda|\pi)]^2 + [1 - \Pr(X \geq \lambda|\pi)]^2, \tag{3}$$

with

$$\Pr(X \geq \lambda|\pi) = \sum_{x=\lambda}^{k} \Pr(X = x|\pi). \tag{4}$$

The right side of Equation 4 can be obtained from the incomplete beta distribution $I_\pi(\lambda, k - \lambda + 1)$, which may simplify computations.[2]

Often, the chance-corrected agreement statistic κ is considered as well, namely,

$$\kappa = \frac{P - P_c}{1 - P_c}, \tag{5}$$

[1] Note that no examinee subscript is used in this report; other descriptions of the procedures discussed here sometimes use notation such as π_j to designate π for examinee j.

[2] This use of the incomplete beta distribution has nothing to do with assuming that proportion-correct true scores have a beta distribution.

where

$$P_c = [\Pr(X \geq \lambda)]^2 + [1 - \Pr(X \geq \lambda)]^2. \tag{6}$$

The first term in Equation 6 is the squared proportion of examinees who pass, and the second term is the squared proportion of examinees who fail.

The principal challenge in using Subkoviak's procedure is to estimate p when the examinee's proportion-correct true score, π, is unknown. Subkoviak makes two suggestions. First, as an estimate of π use the examinee's observed proportion correct score, which we sometimes designate $\hat{\pi}_o$; i.e.,

$$\hat{\pi}_o = \frac{x}{k} = \overline{x}. \tag{7}$$

Second, use Kelley's (1947) regressed-score estimate as an estimate of π, namely,

$$\hat{\pi}_r = (1 - \rho^2)\,\mu + \rho^2\,\overline{x}, \tag{8}$$

where ρ^2 is reliability, and μ is the grand mean—i.e., the expected value (over examinees) of $\overline{x}$. (If $n < \infty$, then the grand mean is $\sum \overline{x}/n$, where the sum is taken over examinees.)

The principal purpose of this paper is to examine the amount of bias in $\hat{P}$ and $\hat{\kappa}$ that results from applying these two estimates of true score, as well as two other estimates that are considered. In doing so, we propose an "optimally" weighted estimate of π that appears to reduce bias in $\hat{P}$ and $\hat{\kappa}$ substantially. We focus principally on the dichotomous-data case, but we also consider extensions to polytomous data and to complex assessments.

2 Dichotomous Data and the Binomial Error Model

Suppose the data were generated such that they conform perfectly to the beta-binomial model. Under these circumstances, Huynh's results for P, P_c, and κ are the parameter values, and we know that the observed scores have a negative hypergeometric distribution, $f(X) \sim NH(a, b)$. To study Subkoviak's suggested use of $\overline{x}$ we proceed as follows:

1. for each possible x ($x = 0, 1, 2, \ldots, k$), obtain the result in Equation 4 with $\overline{x}$ replacing π;
2. use Equation 3 to estimate the $k + 1$ estimates of p;
3. multiply the $k+1$ estimates obtained in Step 2 by the corresponding $f(x)$;
4. sum the results in Step 2 to get $\hat{P}$.

To study Kelley's regressed-score estimate we follow the same steps except that $\hat{\pi}_r$ in Equation 8 replaces π in Step 1.

To obtain $\hat{P}_c$ we follow similar steps. Specifically, using examinee observed mean scores as estimates of π, the steps are:

1. for each possible x ($x = 0, 1, 2, \ldots, k$), obtain the result in Equation 4 with $\overline{x}$ replacing π;
2. multiply the $k+1$ estimates obtained in Step 1 by the corresponding $f(x)$;
3. sum the results in Step 2 to get an estimate of $\Pr(X \geq \lambda)$;
4. use Equation 6 to get $\hat{P}_c$.

To get $\hat{P}_c$ based on Kelley's regressed-score estimates we follow the same steps except that $\hat{\pi}_r$ in Equation 8 replaces π in Step 1. Given $\hat{P}$ and $\hat{P}_c$, obviously κ can be estimated.

2.1 Simulations

The procedures outlined above were used for two beta-binomial simulations: (1) $a = 4$, $b = 4$, and $k = 20$; and (2) $a = 8$, $b = 4$, and $k = 20$. The true-score and observed-score PDFs for these two simulation conditions are provided in Figures 1 and 2, respectively. Obviously, the distributions are symmetric for the first simulation, while they are negatively skewed for the second simulation.

Specifically, for these simulations it is assumed that X and X' are two randomly parallel forms of a test with k dichotomously-scored items, the regression of true scores on observed scores is linear, and true scores are distributed as beta with parameters a and b. Under these assumptions, the marginal distribution of X (or X') is negative hypergeometric, the joint distribution of X and X' is bivariate negative hypergeometic with a correlation of

$$\text{KR21} = \frac{k}{k + (a + b)},$$

and the probability of agreement for the group of examinees is

$$\Pr[(X < \lambda) \text{ and } (X' < \lambda)] + \Pr[(X \geq \lambda) \text{ and } (X' \geq \lambda)].$$

This is the parameter P for a beta-binomial with parameters a and b.[3] The parameter P_c is given by Equation 6, and the parameter κ is given by Equation 5.

The results for the two simulations, for several different cut scores, are provided in Tables 1 and 2, where KR21 is denoted ρ^2_{21}. The first quarter of each table provides results based on using observed scores as estimates of true scores; the second quarter provides results based on using regressed-score estimates. Three facts are immediately obvious. First, there is considerable bias in $\hat{P}$, $\hat{P}_c$, and $\hat{\kappa}$, for both observed scores and regressed-score estimates. Second, for these statistics with a given cut score, the absolute value of the bias tends to be about the same for observed scores and regressed-score estimates. Third, the direction of the bias tends to be reversed for observed scores and regressed-score estimates. That is, if one estimate is positively biased, then the corresponding estimate is likely to be negatively biased.

[3] For more details about the beta-binomial see Huynh (1976) and Lord and Novick (1968, esp., p. 517).

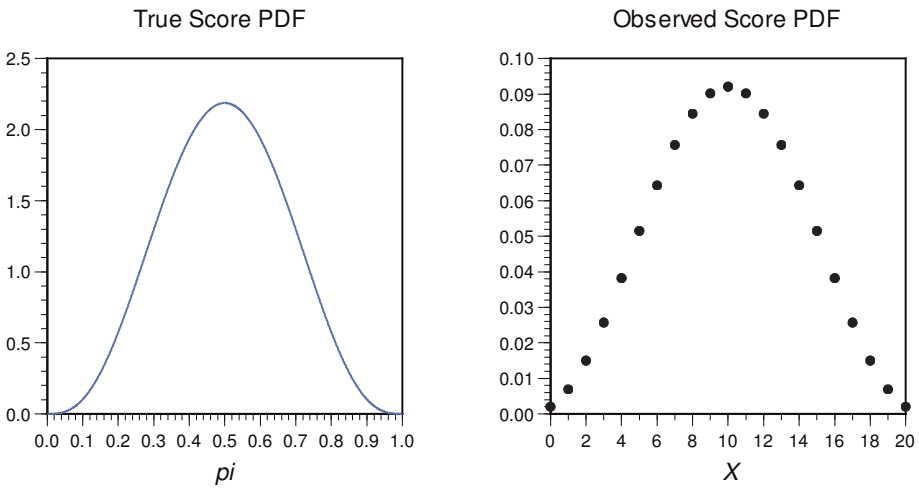

Figure 1: True-score and observed-score PDFs for $a = 4$, $b = 4$, and $k = 20$.

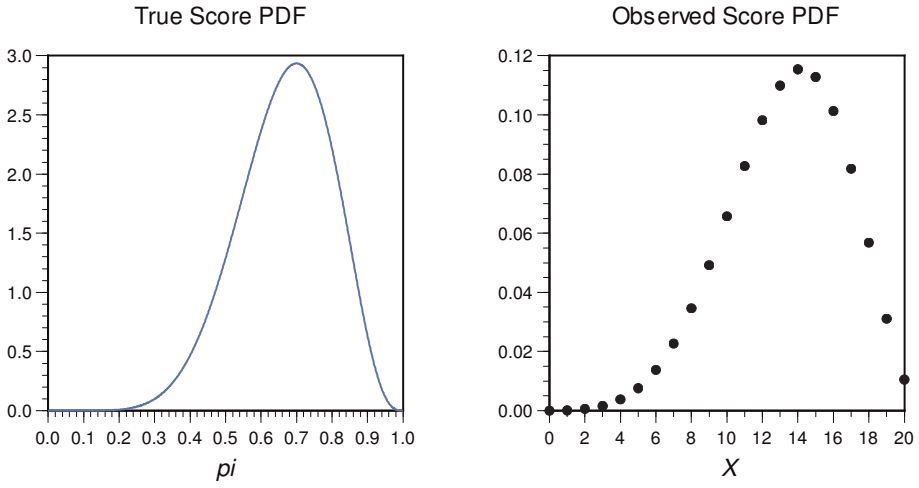

Figure 2: True-score and observed-score PDFs for $a = 8$, $b = 4$, and $k = 20$.

Table 1: Results for $a = 4$, $b = 4$, and $k = 20$ ($\rho^2_{21} = .714$)

	Agreement			Chance Agreement			kappa		
Cut	P	$\hat{P}$	*bias*	P_c	$\hat{P}_c$	*bias*	κ	$\hat{\kappa}$	*bias*
				Using Observed Scores					
3	.966	.952	-.014	.953	.914	-.039	.277	.443	.166
6	.867	.868	.001	.760	.713	-.047	.444	.539	.095
9	.775	.802	.027	.537	.528	-.009	.514	.581	.067
12	.775	.802	.027	.537	.528	-.009	.514	.581	.067
15	.867	.868	.001	.760	.713	-.047	.444	.539	.095
18	.966	.952	-.014	.953	.914	-.039	.277	.443	.166
				Using Regressed Score Estimates					
3	.966	.980	.013	.953	.977	.023	.277	.131	-.146
6	.867	.872	.005	.760	.805	.045	.444	.342	-.102
9	.775	.749	-.027	.537	.547	.010	.514	.445	-.069
12	.775	.749	-.027	.537	.547	.010	.514	.445	-.069
15	.867	.872	.005	.760	.805	.045	.444	.342	-.102
18	.966	.980	.013	.953	.977	.023	.277	.131	-.146
				Using Mean of Observed Scores and Regressed Score Estimates					
3	.993	.993	-.001	.992	.991	-.001	.145	.150	.005
6	.944	.942	-.002	.918	.913	-.004	.316	.327	.012
9	.834	.835	.001	.709	.704	-.005	.432	.442	.011
12	.744	.749	.005	.509	.509	-.000	.478	.488	.009
15	.783	.785	.002	.609	.606	-.004	.445	.455	.010
18	.926	.923	-.003	.894	.889	-.005	.297	.306	.009
				Using Optimal Weight $w = .542$					
3	.966	.966	-.000	.953	.954	.000	.277	.268	-.009
6	.867	.866	-.000	.760	.760	-.000	.444	.443	-.000
9	.775	.776	.000	.537	.537	-.000	.514	.515	.001
12	.775	.776	.000	.537	.537	-.000	.514	.515	.001
15	.867	.866	-.000	.760	.760	-.000	.444	.443	-.000
18	.966	.966	-.000	.953	.954	.000	.277	.268	-.009

Table 2: Results for $a = 8$, $b = 4$, and $k = 20$ ($\rho^2_{21} = .625$)

	Agreement			Chance Agreement			kappa		
Cut	P	$\hat{P}$	*bias*	P_c	$\hat{P}_c$	*bias*	κ	$\hat{\kappa}$	*bias*
				Using Observed Scores					
3	.999	.995	-.004	.999	.993	-.006	.074	.250	.176
6	.979	.963	-.016	.973	.941	-.032	.213	.374	.161
9	.899	.888	-.011	.845	.789	-.055	.347	.469	.122
12	.769	.798	.029	.595	.574	-.020	.430	.525	.095
15	.729	.773	.044	.522	.513	-.009	.432	.533	.101
18	.877	.868	-.009	.823	.748	-.075	.309	.477	.167
				Using Regressed Score Estimates					
3	.999	1.000	.001	.999	1.000	.001	.074	.011	-.062
6	.979	.989	.011	.973	.988	.015	.213	.092	-.121
9	.899	.915	.016	.845	.889	.044	.347	.232	-.115
12	.769	.745	-.024	.595	.616	.021	.430	.336	-.094
15	.729	.688	-.041	.522	.533	.011	.432	.331	-.101
18	.877	.899	.021	.823	.879	.056	.309	.163	-.146
				Using Mean of Observed Scores and Regressed Score Estimates					
3	.999	.998	-.000	.999	.998	-.000	.074	.079	.005
6	.979	.977	-.002	.973	.970	-.003	.213	.227	.014
9	.899	.897	-.002	.845	.838	-.006	.347	.362	.015
12	.769	.773	.004	.595	.592	-.002	.430	.443	.013
15	.729	.734	.005	.522	.521	-.001	.432	.445	.013
18	.877	.875	-.003	.823	.815	-.008	.309	.322	.013
				Using Optimal Weight $w = .558$					
3	.999	.999	.000	.999	.999	.000	.074	.065	-.008
6	.979	.979	-.000	.973	.973	-.000	.213	.210	-.003
9	.899	.898	-.000	.845	.844	-.000	.347	.348	.001
12	.769	.770	.001	.595	.595	-.000	.430	.432	.002
15	.729	.729	.000	.522	.522	-.000	.432	.433	.001
18	.877	.876	-.001	.823	.823	.000	.309	.302	-.007

This suggests that perhaps a better estimate of π for an examinee would be to weight them equally—i.e., to obtain the mean of the two estimates:

$$\hat{\pi}_e = .5\,[(1-\rho^2)\,\mu + \rho^2\,\overline{x}] + .5\,\overline{x}, \tag{9}$$

where the prescript e stands for equal weighting. Results for this estimate are provided in the third quarter of each of the tables. Clearly using $\hat{\pi}_e$ as an estimate of π works quite well.[4]

2.2 An Optimal Estimate of π

Even though using the mean of the examinee's observed score and regressed-score estimate appears to work quite well, it is an ad hoc solution. A more satisfying and perhaps better solution would be to weight the two components such that the resulting weighted estimate had some desirable property. Recall that ideally we would like to get each examinee's agreement statistic conditional on the examinee's true score (see Equations 3 and 4). This suggests that we might want to choose weights such that the variance (over examinees) of the resulting weighted estimate is true score variance, which we can estimate if we have an estimate of reliability.

Let w be the weight for the regressed-score estimate. Then

$$\hat{\pi}_w = w\,[(1-\rho^2)\,\mu + \rho^2\,\overline{x}] + (1-w)\,\overline{x}. \tag{10}$$

We want

$$\mathrm{Var}(\hat{\pi}_w) = \sigma^2_{(T/k)} = \rho^2\sigma^2_{(X/k)}, \tag{11}$$

where ρ^2 is reliability, and $\sigma^2_{(T/k)}$ and $\sigma^2_{(X/k)}$ are true-score and observed-score variance, respectively, in the proportion-correct metric. Now,

$$\begin{aligned}
\mathrm{Var}(\hat{\pi}_w) &= \mathrm{Var}\{w[(1-\rho^2)\,\mu + \rho^2\,\overline{x}] + (1-w)\,\overline{x}\} \\
&= \mathrm{Var}[w(1-\rho^2)\,\mu + (w\,\rho^2 + 1 - w)\,\overline{x}] \\
&= \mathrm{Var}[(w\,\rho^2 + 1 - w)\,\overline{x}] \\
&= (w\,\rho^2 + 1 - w)^2\,\sigma^2_{(X/k)}.
\end{aligned} \tag{12}$$

Equating Equations 11 and 12 gives

$$\begin{aligned}
\rho^2 &= (w\,\rho^2 + 1 - w)^2 \\
&= [w(\rho^2 - 1) + 1]^2,
\end{aligned} \tag{13}$$

[4] Algina and Noe (1978) reported simulations in which they too found that using either observed mean scores or regressed-scores as estimates of π led to biased estimates of P (they did not study κ), with different directions for the bias. They even suggested using an estimate like $\hat{\pi}_e$ in Equation 9 (Algina & Noe, p. 109), but they did not study it in their simulations. One difference between their simulations and the simulations in this paper is that Algina and Noe use KR20, rather than KR21, to obtain regressed-score estimates in their simulations.

which can be expanded, simplified, and solved for w using the usual formula for the solution of a quadratic equation. However, in this case it is clear by inspection that the solution is

$$\begin{aligned} w &= \frac{\rho - 1}{\rho^2 - 1} \\ &= \frac{\rho - 1}{(\rho + 1)(\rho - 1)} \\ &= \frac{1}{\rho + 1}, \end{aligned} \tag{14}$$

where it should be noted that ρ is the *square-root* of reliability (sometimes called the index of reliability in older literature). In short, using $w = 1/(\rho + 1)$ as the weight for the regressed-score estimate (and $1 - w$ as the weight for the observed mean score) makes the variance of $\hat{\pi}_w$ equal to true score variance in the proportion-correct metric. In this sense, $w = 1/(\rho + 1)$ is the "optimal" weight. Note that as ρ^2 goes from 0 to 1, w goes from 1 to .5. For example,

ρ^2:	.1	.2	.3	.4	.5	.6	.7	.8	.9
w:	.76	.69	.65	.61	.59	.56	.54	.53	.51

Given $w = 1/(\rho + 1)$, and using the notation and simplifications that led to Equation 14, a simpler expression for $\hat{\pi}_w$ in Equation 10 can be derived:

$$\begin{aligned} \hat{\pi}_w &= \frac{(1 - \rho^2)\,\mu + \rho^2\,\overline{x}}{\rho + 1} + \frac{\rho}{\rho + 1}\,\overline{x} \\ &= (1 - \rho)\,\mu + \left(\frac{\rho^2 + \rho}{\rho + 1}\right)\,\overline{x} \\ &= (1 - \rho)\,\mu + \rho\,\overline{x}. \end{aligned} \tag{15}$$

Clearly, Equation 15 has the form of a regressed-score estimate, but in Equation 15 the slope is the square-root of reliablilty as opposed to reliability itself.

2.3 Simulations Revisited

The bottom quarter of Tables 1 and 2 provides results for the two simulations using optimal weights. It is evident that using optimal weights is almost always better than using equal weights in these simulations. In fact, to three decimal digits, the absolute value of the bias in $\hat{P}$ and $\hat{P}_c$ is 0.000 for the symmetric case and almost always 0.000 for the skewed case. There is clearly more bias in $\hat{\kappa}$ than in $\hat{P}$ or $\hat{P}_c$, no matter what estimate is used, but the amount of bias in $\hat{\kappa}$ using $\hat{\pi}_w$ appears tolerable (less than .01).

With minor exceptions, the estimates of κ are considerably less biased using optimal weights than using equal weights. The exceptions occur primarily for the very low cut score of 3 items correct (out of $k = 20$). This last case should not be taken too seriously, however, because there is only a very small proportion of the examinees in this region of the score scale, especially for the skewed case

in Table 2 (see Figure 1). With very small proportions of examinees above or below a cut score, the magnitudes of both $\hat{P}$ and $\hat{P}_c$ will be very high, and $\hat{\kappa}$ will be very low and relatively unstable in the sense that a small change in $\hat{P}_c$ can cause a considerable change in $\hat{\kappa}$.

Nothing in the derivation of the optimal weight $w = 1/(\rho + 1)$ involves the skewness (or other higher-order moments) of the observed score distribution. Therefore, one might speculate that $\hat{\pi}_w$ might not work as well for skewed observed score distributions as it does for symmetric ones. The simulation results are inconsistent with this speculation, however. It appears to the authors that the skewness of the observed score distribution per se is not a crucial issue in how well $\hat{\pi}_w$ works; rather, other things being equal, it appears that $\hat{\pi}_w$ generally works well when the observed and true score distributions have *similar* degrees of skewness. (In Figure 2 the skewnesses of the true-score and observed-score distributions are $-.364$ and $-.375$, respectively.)

Compared to equal weighting, optimal weighting tends to make a somewhat greater improvement in κ in the skewed case with $w = .558$ and $\rho^2_{21} = .625$, than in the symmetric case with $w = .542$ and $\rho^2_{21} = .714$. This is consistent with the fact that the optimal weight for the skewed case is somewhat further from .5 than is the optimal weight for the symmetric case. That is, relative to equal weighting, optimal weighting has a somewhat larger influence in the skewed case, which has the smaller reliability.

2.4 Parametric Bootstrap

The simulation results in Tables 1 and 2 were obtained using Equation 4 directly. More specifically, $\Pr(X \geq \lambda|\pi)$ was obtained using the incomplete beta distribution $I_\pi(\lambda, n - \lambda + 1)$ with π replaced by one of the four estimates discussed previously. An alternative procedure would be to use the parametric bootstrap described in Brennan and Wan (2004). Specifically, in theory $\Pr(X \geq \lambda|\pi)$ can be obtained as follows:

1. Set $N = 0$.
2. Draw a uniform random number, say u.
3. If $u \leq \pi$, set the item response to 1; if $u > \pi$, set the item response to 0.
4. Repeat steps 2 and 3 k times, and let the number of 1's be x.
5. If $x \geq \lambda$ increment N by 1.
6. Repeat steps 2–5 B times.
7. Compute N/B.

If $B \to \infty$, $\Pr(X \geq \lambda|\pi) = N/B$ in step 7. This is called the parametric bootstrap with an infinite number of replications.

To use the parametric bootstrap for estimation, we simply replace π in step 3 with one of the four estimates in Equations 7–10, and choose a finite value

for B ($B \geq 1000$ seems desirable). Brennan and Wan (2004) discuss doing so using examinee observed mean scores and regressed-score estimates. This paper suggests that these two estimates are likely to result in considerable bias for group-level statistics. For such statistics, it appears better to use $\hat{\pi}_e$ or $\hat{\pi}_w$, and preferably the latter.

2.5 Estimating Reliability

Obviously, a reliability coefficient (or its square root) must be chosen to apply the methodology discussed here. Huynh (1976) uses KR21, based on the fact that Lord and Novick (1968, p. 523) have shown that for the beta-binominal model, reliability (in the sense of the squared correlation between true and observed scores) is KR21. For these reasons, the simulation results in Tables 1 and 2 use KR21, which is denoted ρ^2_{21}.

In decision consistency contexts, it is the absolute magnitude of an examinee's score that is of interest, rather than the examinee's score relative to others. Therefore, in choosing a reliability coefficient, the crucial issue is that the error variance be absolute error variance, $\sigma^2(\Delta)$, using the terminology and notation in generalizability theory (see Brennan, 2001). Brennan (2001) and Brennan and Lee (2006) have shown that $\sigma^2(\Delta)$ is incorporated in both ρ^2_{21} and Φ, which tend to have similar estimated values with real data.[5] So, Φ might be used[6] instead of ρ^2_{21}. By contrast, KR20 cannot be recommended because it incorporates relative error variance, $\sigma^2(\delta)$, not $\sigma^2(\Delta)$ (see Brennan, 2001).

2.6 Other Issues

The simulations demonstrate the $\hat{\pi}_w$ works well (at least for the studied conditions) when the beta-binomial model holds. By contrast, strictly speaking these simulations do not demonstrate how well $\hat{\pi}_w$ works when the data do not conform to the beta-binomial model. Still, it seems almost certain that $\hat{\pi}_w$ will work much better than using the examinee observed mean score or the regressed-score estimate because (a) conditioning on $\hat{\pi}_w$ is closer to conditioning on π, and (b) the variance of the $\hat{\pi}_w$ is true-score variance. Note that

$$\sigma^2_{\hat{\pi}_r} < \sigma^2_{\hat{\pi}_w} = \sigma^2_{(T/k)} < \sigma^2_{(X/k)}.$$

3 Polytomous Data and the Multinomial Error Model

Suppose a test (or test section) consists of k polytomous items, and each item is scored as one of h possible score points, $c_1 < c_2 < \cdots < c_h$. Assume a sample of

[5] Φ is usually called an index of dependability to distinguish it from a generalizability coefficient $\boldsymbol{E}\rho^2$. The principal difference is that Φ uses absolute error variance, $\sigma^2(\Delta)$, while $\boldsymbol{E}\rho^2$ uses relative error variance, $\sigma^2(\delta)$.

[6] If all persons take the same items, then Φ for the $p \times I$ design should be used; if all persons take different items then Φ for the $I{:}p$ design should be used.

k items is drawn at random from a universe of items, and let $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_h\}$ denote the proportions of items in the universe for which an examinee would get scores of $c_1, c_2, \ldots, c_h$, respectively. Further, let $X_1, X_2, \ldots, X_h$ be random variables representing the number of items scored $c_1, c_2, \ldots, c_h$, respectively, such that $X_1 + X_2 + \cdots + X_h = k$. It follows that $X = c_1X_1 + c_2X_2 + \cdots + c_hX_h$ is the total raw score. Note that $X_1, X_2, \ldots, X_h$ are random variables that have a multinomial distribution:

$$\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_h = x_h, |\boldsymbol{\pi}) = \frac{k!}{x_1!x_2! \cdots x_h!} \, \pi_1^{x_1} \, \pi_2^{x_2} \cdots \pi_h^{x_h}. \quad (16)$$

This description of the multinomial model mirrors that provided by Lee (2005a, b), Brennan and Wan (2004), and Lee, Wang, Kim, and Brennan (2006).

Equation 16 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, where the score categories are simply $c_1 = 0$ and $c_2 = 1$.

There are a number of sets of values for $x_1, x_2, \ldots, x_h$ that give a particular $X = x$. So, in general, using Equation 16, the probability of a particular $X = x$ score is:

$$\Pr(X = x|\boldsymbol{\pi}) = \sum_{c_1x_1 + c_2x_2 + \cdots + c_hx_h = x} \Pr(X_1 = x_1, X_2 = x_2, \ldots, X_h = x_h, |\boldsymbol{\pi}), \quad (17)$$

where the sum is taken over all values of $c_1\,x_1, c_2\,x_2, \ldots, c_h\,x_h$ that sum to x. It follows that the probability of passing is:

$$\Pr(X \geq \lambda|\boldsymbol{\pi}) = \sum_{x=\lambda}^{\max(x)} \Pr(X = x|\boldsymbol{\pi}), \quad (18)$$

where $\Pr(X = x|\boldsymbol{\pi})$ is given by Equation 17. Equation 18 plays the same role as Equation 4 in the series of Equations 2–6 for obtaining decision consistency indexes.

3.1 Estimating ${}_w\hat{\pi}_l$

When $\boldsymbol{\pi}$ is not known, it is natural to consider using $\hat{\pi}_1 = \overline{x}_1$, $\hat{\pi}_2 = \overline{x}_2$, ..., $\hat{\pi}_h = \overline{x}_h$, where $\overline{x}_l$ $(l = 1, 2, \ldots, h)$ is the proportion of times (over the k items) that the examinee's score was c_l. This is analogous to using the examinee's observed mean score in the binomial case. We denote this estimator as

$${}_o\hat{\pi}_l = \overline{x}_l. \quad (19)$$

A multinomial version of Kelley's regressed-score estimate is

$${}_r\hat{\pi}_l = (1 - \rho^2)\,\mu_l + \rho^2\,\overline{x}_l, \quad (20)$$

where μ_l is the expected value over examinees of $\overline{x}_l$. A third possible estimate is obtained by equally weighting ${}_r\hat{\pi}_l$ and ${}_o\hat{\pi}_l = \overline{x}_l$:

$$ {}_e\hat{\pi}_l = .5\,({}_r\hat{\pi}_l) + .5\,(\overline{x}_l). \tag{21}$$

A fourth possible estimate is to optimally weight ${}_r\hat{\pi}_l$ and $\overline{x}_l$ in the manner discussed in section 2.2. As shown in the derivation that produced Equation 15, the optimally weighted estimate is equivalent to

$$ {}_w\hat{\pi}_l = (1-\rho)\,\mu_l + \rho\,\overline{x}_l. \tag{22}$$

For reasons discussed in the next section, for the last three estimates it is suggested that reliability (or its square root) be of the Φ type (see Brennan, 2001), or something close to it, for the full-length k-item test. Although logic and the simulation results discussed in section 2 stongly suggest that ${}_w\hat{\pi}_l$ is the preferable estimate for group-level decision consistency estimates, strictly speaking a more compelling argument would require additional simulations tailored to polytomous data and the multinomial model.

3.2 Properties of ${}_w\hat{\pi}_l$

Note that

$$ \sum_{l=1}^{h} {}_w\hat{\pi}_l = (1-\rho)\sum_{l=1}^{h}\mu_l + \rho\sum_{l=1}^{h}\overline{x}_l. \tag{23}$$

Recall that $\overline{x}_l$ is the proportion of times (over the k items) that the examinee's score is c_l. Clearly, for every examinee the sum of these proportions over the h categories is 1, which means that the sum of the μ_l is also 1. Therefore,

$$ \sum_{l=1}^{h} {}_w\hat{\pi}_l = (1-\rho) + \rho = 1, \tag{24}$$

as must be the case for the multinomial model to hold.[7]

Recall that the motivation behind developing the optimally weighted estimate of true score was that the variance of the estimates (over examinees) be true score variance. We show next that this characteristic also applies to the k-item mean score based on the ${}_w\hat{\pi}_l$. Without loss of generality, suppose there are $h = 3$ score categories. Then, let ${}_w\hat{\pi}$ be the optimally weighted estimate of mean score (over all k items) in the sense that

$$\begin{aligned} {}_w\hat{\pi} &= c_1({}_w\hat{\pi}_1) + c_2({}_w\hat{\pi}_2) + c_3({}_w\hat{\pi}_3) \\ &= [c_1(1-\rho)\mu_1 + c_2(1-\rho)\mu_2 + c_3(1-\rho)\mu_3] \\ &\quad + \rho\,[c_1\,\overline{x}_1 + c_2\,\overline{x}_2 + c_3\,\overline{x}_3]. \end{aligned}$$

[7] A similar proof shows that $\sum_{l=1}^{h} {}_r\hat{\pi}_l = 1$, as well.

It follows that

$$\begin{aligned} \text{Var}({}_w\hat{\pi}) &= \rho^2 \,\text{Var}[c_1\,\overline{x}_1 + c_2\,\overline{x}_2 + c_3\,\overline{x}_3] \\ &= \rho^2 \sigma^2_{(X/k)} \\ &= \sigma^2_{(T/k)}, \end{aligned} \tag{25}$$

where $\sigma^2_{(T/k)}$ and $\sigma^2_{(X/k)}$ are the true-score and observed-score variance, respectively, in the mean-score metric.

The derivation of Equation 25 clearly requires that the same value for ρ be used for each ${}_w\hat{\pi}_l$, with this value being the square-root of reliability for the full-length test of k items. (Similarly, in the previous section it was suggested that the same value of ρ^2 be used for the h regressed-score estimates.) This use of reliability for the full-length test is consistent with the principal goal here, namely, to get an estimate of π for the full-length test that is close to π itself.

Still, using a full-length estimate of reliability may appear strange since we are relating ${}_w\hat{\pi}_l$ and $\overline{x}_l$ for each individual category, which seems to suggest using category-specific ρ's, which we designate generically as ρ_l. However, in this case, "category" is not a set of items; it is a score category, and the examinee observed mean score for category l is the proportion of times that the examinee's responses to the k items were scored c_l. In theory, ρ_l could be estimated for $\overline{x}_l$, but it would not be particularly meaningful for present purposes. For example, the sum of the h error variances associated with these values of ρ_l would not be the error variance for the total score, because the category error scores are correlated (recall that the $\overline{x}_l$ sum to 1 for each examinee).

3.3 Parametric Bootstrap

Section 2.4 discussed the parametric bootstrap for the binomial. The steps discussed there apply as well to the multinomial with steps 3 and 4 replaced by:

3. If $0 \le u < \pi_1$, set the item response to c_1; if $\pi_1 \le u < (\pi_1 + \pi_2)$, set the item response to c_2; ...; if $\sum_{l=1}^{h-1} \pi_l \le u \le 1$, set the item response to c_h.

4. Repeat steps 2 and 3 k times to get the number of scores associated with each category, x_l, and determine $x = c_1\,x_1 + c_2\,x_2 + \cdots + c_h\,x_h$.

To use the parametric bootstrap for estimation, we simply replace the π_l in step 3 with one of the four estimates in Equations 19–22, and choose a finite value for B ($B \ge 1000$ seems desirable).

3.4 Estimating Reliability

For decision consistency indexes, interest focuses on the absolute value of examinee scores, not simply a rank ordering of these scores. Consequently, as noted in section 2.5, the appropriate error variance is absolute error variance, $\sigma^2(\Delta)$, which should be the error variance incorporated in the k-item reliability-like coefficient (or its square root) chosen for use in ${}_r\hat{\pi}_l$, ${}_e\hat{\pi}_l$, or ${}_w\hat{\pi}_l$. Consequently,

Φ for the k-item test is an appropriate coefficient.[8] Alternatively, a coefficient discussed by Lee (2005a, Equation 10) might be used.

4 Complex Assessments and the Compound Multinomial Error Model

It is logically straightforward, but mathematically complex, to extend the methodology discussed in this paper to complex assessments involving multiple sets of items that differ in some way (e.g., different content categories and/or sets of items with different numbers of score categories). The crucial step is to model errors in such complex assessments according to the compound multinomial model, as discussed in considerable detail by Lee (2005a, b). Here we merely outline the relatively simple case of a test that consists of k_1 dichotomous items and k_2 polytomous items each of which has three categories, with the total score defined as the sum of the scores on the two types of items.

Errors for the polytomous items are modeled by the multinomial. Errors for the dichotomous items are modeled by the binomial, which is a special case of the multinomial with two categories having $c_1 = 0$ and $c_2 = 1$. Therefore, from the multinomial perspective, the total score on the dichotomous items is $x = c_1\, x_1 + c_2\, x_2 = x_2$, where x_2 is literally the number of items with a score of $c_2 = 1$, which is obviously the total score x.

Here, for the polytomous items we will use the same notational convention used previously. However, to keep notation distinct for the dichotomous and polytomous sets of items (without introducing multiple subscripts), for the dichotomous items we will use Y as the total score random variable and τ as the proportion-correct true score. Then, the total-score random variable Z over both dichotomous and polytomous items is

$$Z = Y + X = Y + (c_1 X_1 + c_2 X_2 + c_3 X_3).$$

Assuming independence of errors (given τ and $\boldsymbol{\pi}$)

$$\Pr(Z = z|\tau, \boldsymbol{\pi}) = \sum \Pr(Y = y|\tau) \Pr(X_1 = x_1, X_2 = x_2, \ldots, X_h = x_h, |\boldsymbol{\pi}), \tag{26}$$

where the sum is taken over all values of $y, c_1\, x_1, c_2\, x_2, \ldots, c_h\, x_h$ that sum to z; $\Pr(Y = y|\tau)$ is given by the binomial PDF in Equation 1 (with the obvious change of Y to X and y to x); and $\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_h = x_h, |\boldsymbol{\pi})$ is given by the multinomial PDF in Equation 16. It follows that the probability of passing is:

$$\Pr(Z \geq \lambda|\tau, \boldsymbol{\pi}) = \sum_{z=\lambda}^{\max(z)} \Pr(Z = z|\tau, \boldsymbol{\pi}), \tag{27}$$

[8] If all persons take the same items, then Φ for the $p \times I$ design should be used; if all persons take different items then Φ for the $I{:}p$ design should be used.

where $\Pr(Z = z|\tau, \boldsymbol{\pi})$ is given by Equation 26. Equation 27 plays the same role as Equation 4 in the series of Equations 2–6 for obtaining decision consistency indexes, with the obvious change of Z to X.

4.1 Estimation

For estimation, we have the following four possible sets of substitutes for τ and π_l:

1. use the observed mean scores in Equations 7 and 19;
2. use the regressed-score estimates in Equations 8 and 20;
3. use the averages of observed mean scores and regressed-score estimates in Equations 9 and 21; and
4. use the optimally weighted estimates in Equations 10 and 22.

The last alternative is likely to be the best.

For the same types of reasons as discussed in sections 3.1–3.4, it is suggested that reliability (or its square root) involve absolute error variance for the full-length test with $k = k_1 + k_2$ items. For the example considered here, an appropriate coefficient would be the multivariate Φ dependability index with two fixed categories—dichotomously-scored items and polytomously-scored items (see Brennan, 2001, chap. 10, for details). Alternatively, a coefficient discussed by Lee (2005a, section 3.1) might be used.

4.2 Parametric Bootstrap

Sections 2.4 and 3.3 considered parametric bootstrap procedures for the binomial and multinomial error models, respectively. For the compound multinomial example considered here the parametric bootstrap is the conjunction of the two, and the steps are:

1. Set $N = 0$.
2. Draw a uniform random number, say u.
3. If $u \leq \tau$, set the dichotomous-item response to 1; if $u > \tau$, set the response to 0.
4. Repeat steps 2 and 3 k_1 times, and let the number of 1's be y.
5. Draw a uniform random number, say u.
6. If $0 \leq u < \pi_1$, set the item response to c_1; if $\pi_1 \leq u < (\pi_1 + \pi_2)$, set the item response to c_2; ...; if $\sum_{l=1}^{h-1} \pi_l \leq u \leq 1$, set the item response to c_h.
7. Repeat steps 5 and 6 k_2 times to get the number of scores associated with each category, x_l, and determine $x = c_1\, x_1 + c_2\, x_2 + \cdots + c_h\, x_h$.

8. Set $z = y + x$; if $z \geq \lambda$ increment N by 1.

9. Repeat steps 2–8 B times.

10. Compute N/B.

If $B \rightarrow \infty$, $\Pr(Z \geq \lambda|\tau, \boldsymbol{\pi}) = N/B$ in step 10. This is called the parametric bootstrap with an infinite number of replications. To use the parametric bootstrap for estimation, we simply replace τ in step 3 and the π_l in step 6 with one of the four pairs of estimates discussed in section 4.1, and choose a finite value for B ($B \geq 1000$ seems desirable).

5 Concluding Comments

The principal results of this paper are the equations that define the optimal estimate of π for an examinee (Equations 10, 14, and 15 for the binomial error model; and Equation 22 for the multinomial error model). Logic suggests that using the optimal estimate should lead to less bias in P and κ than using observed mean scores, regressed-score estimates, or the average of the two. This logic is supported by the simulations for dichotomous data provided in this paper. Of course, other dichotomous-data simulations could be conducted to further examine optimal estimates, and simulations involving the multinomial and compound multinomial model have yet to be conducted.

The optimal estimate is "optimal" in the sense discussed in section 2.2 for group-level statistics. It is not optimal in the least-squares sense used in regression. Also, it is important to note that for examinee-level measures of agreement such as p in Equation 3, estimating examinee true score using the examinee's observed mean score has the distinct advantage of providing an unbiased estimate. This is an example of the well-known fact that a statistic that works well (in some sense) for estimating an examinee-level result does not necessarily work well at the group level, and vice-versa.

For each of the three models discussed here (binomial, multinomial, and compound multinomial) a corresponding parametric bootstrap procedure has been discussed (see also Brennan & Wan, 2004). When the computation of examinee scores is not very complex, there is no particular advantage to the bootstrap over the model-based procedures. However, the bootstrap has the distinct advantage of being quick,[9] easy to implement, and very flexible in the sense that it can accommodate very complex scoring, equating, weighting, and/or scaling procedures. Recall that the parametric bootstrap leads to a vector of item scores for as many strata (content or format) as there may be. Once these vectors are available, scoring, equating, weighting, and/or scaling of any kind can be done in the same manner as with the original data.

[9]With even a modestly large number of strata, the compound multinomial can require a considerable amount of computational time.

6 References

Algina, J., & Noe, M. J. (1978). A study of the accuracy of Subkoviak's single-administration estimate of the coefficient of agreement using two true-score estimates. *Journal of Educational Measurement, 15*, 101–110.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.

Brennan, R. L., & Lee, W. (2006, March). *Some perspectives on KR-21.* (CASMA Technical Note No 2). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available from http://www.education.uiowa.edu/casma)

Brennan, R. L., & Wan, L. (2004, June). *A bootstrap procedure for estimating decision consistency for single-administration complex assessments.* (CASMA Research Report No 7). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available from http://www.education.uiowa.edu/casma)

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13*, 253–264.

Kelley, T. L. (1947). *Fundamentals of statistics.* Cambridge: Harvard University Press.

Lee, W. (2005a, January). *A multinomial error model for tests with polytomous items.* (CASMA Research Report No. 10). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma)

Lee, W. (2005b, November). *Classification consistency under the compound multinomial model.* (CASMA Research Report No. 13). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma)

Lee, W., Wang, T., Kim, S., &, Brennan, R. L. (2006, April). *A strong ture-score model for polytomous items.* (CASMA Research Report No. 16). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma)

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13*, 265–276.