Variance over Schools of School True-Score Means for Samples of Size N

Robert L. Brennan
Center for Advanced Studies in Measurement and Assessment
University of Iowa

July 11, 2003

The theorems that follow rely heavily on two relatively unknown but easily proven results. Letting Y be a random variable with realizations $y_1, y_2, \ldots, y_n$,

$$\sigma^2(Y) = \frac{1}{n^2} \sum\sum_{i<j} (y_i - y_j)^2, \tag{1}$$

where $i, j = 1, 2, \ldots, n$. When $n = 2$, a special case of Equation 1 is

$$\sigma^2(Y) = \frac{1}{4}(y_1 - y_2)^2. \tag{2}$$

The proof of Equation 1 is as follows:

$$\begin{aligned}
\frac{1}{n^2} \sum\sum_{i<j} (y_i - y_j)^2 &= \frac{1}{2n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} (y_i - y_j)^2 \\
&= \frac{1}{2n^2} \sum_i \sum_j \left(y_i^2 + y_j^2 - 2\, y_i y_j\right) \\
&= \frac{1}{2n^2} \left[n \sum_i y_i^2 + n \sum_j y_j^2 - 2 \sum_i y_i \sum_j y_j \right] \\
&= \frac{\sum_i y_i^2}{2n} + \frac{\sum_j y_j^2}{2n} - \overline{y}_i \overline{y}_j \\
&= \frac{\sum_i y_i^2}{n} - \overline{y}_i^2 \\
&= \sigma^2(Y).
\end{aligned}$$

The proof of Equation 2 follows immediately from Equation 1.

Theorem: Using the notation in Hill and DePascale (2003), with $n = 2$ schools,

$$\sigma^2(\overline{T}_0) = \frac{\sigma^2(T|S)/N}{2} + \sigma^2(\overline{T}), \tag{3}$$

where $\sigma^2(\overline{T})$ is the variance of school mean true scores for the population of students (assumed to be essentially infinite for both schools), $\sigma^2(T|S)$ is the variance of student true scores within a school (assumed to be the same for both schools), and $\sigma^2(\overline{T}_0)$ is the variance of school mean true scores for samples of size N.

Proof: Let the two schools be labeled A and B. Also, let $y_1 = \overline{T}_{A0}$ be the mean of student true scores for a sample of size N from school A. Similarly, $y_2 = \overline{T}_{B0}$. Then, using Equation 2, the expected value of the variance of $\overline{T}_{A0}$ and $\overline{T}_{B0}$ is:

$$\begin{aligned}
\sigma^2(\overline{T}_0) &= \boldsymbol{E}\left[\frac{1}{4}(\overline{T}_{A0} - \overline{T}_{B0})^2\right] \\
&= \frac{1}{4}\left[\boldsymbol{E}\,\overline{T}^2_{A0} + \boldsymbol{E}\,\overline{T}^2_{B0} - 2\,\boldsymbol{E}\,\overline{T}_{A0}\overline{T}_{B0}\right] \\
&= \frac{1}{4}\Big\{[\boldsymbol{E}\,\overline{T}^2_{A0} - (\boldsymbol{E}\overline{T}_{A0})^2] + [\boldsymbol{E}\,\overline{T}^2_{B0} - (\boldsymbol{E}\overline{T}_{B0})^2] \\
&\quad + [(\boldsymbol{E}\overline{T}_{A0})^2 + (\boldsymbol{E}\overline{T}_{B0})^2 - 2\,\boldsymbol{E}\,\overline{T}_{A0}\overline{T}_{B0}]\Big\} \\
&= \frac{1}{4}\left[\sigma^2(\overline{T}_{A0}) + \sigma^2(\overline{T}_{B0}) + (\boldsymbol{E}\,\overline{T}_{A0} - \boldsymbol{E}\,\overline{T}_{B0})^2\right], \qquad (4)
\end{aligned}$$

because $\boldsymbol{E}\,\overline{T}_{A0}\overline{T}_{B0} = (\boldsymbol{E}\,\overline{T}_{A0})(\boldsymbol{E}\,\overline{T}_{B0})$ since the students are independent for the two schools. Note that $\sigma^2(\overline{T}_{A0})$ is the variance of the mean of student true scores for samples of size N from school A; similarly, $\sigma^2(\overline{T}_{B0})$ is the variance of the mean of student true scores for samples of size N from school B. Assuming they are equal, then in the notation of Hill and DePascale (2003),

$$\sigma^2(\overline{T}_{A0}) = \sigma^2(\overline{T}_{B0}) = \sigma^2(T|S)/N.$$

It follows that Equation 4 is

$$\sigma^2(\overline{T}_0) = \frac{\sigma^2(T|S)/N}{2} + \frac{(\overline{T}_A - \overline{T}_B)^2}{4}. \tag{5}$$

where $\overline{T}_A$ is the population mean for school A and $\overline{T}_B$ is the population mean for school B. In comparing Equation 2 with the second term in Equation 5, it is evident that the second term is the variance (over schools) of the population mean true scores for schools. Therefore,

$$\sigma^2(\overline{T}_0) = \frac{\sigma^2(T|S)/N}{2} + \sigma^2(\overline{T}). \tag{6}$$

QED

Theorem: For any number of schools, n,

$$\sigma^2(\overline{T}_0) = \left(\frac{n-1}{n}\right)\frac{\sigma^2(T|S)}{N} + \sigma^2(\overline{T}). \tag{7}$$

"Proof": The general proof involves considerable notational complexity. Therefore, we derive the result here for $n = 3$ being careful to use n rather than 3 to hint at the generality of the result for any n. When $n = 3$, there are $(n^2 - n)/2 = n(n-1)/2 = 3$ terms in Equation 1:

$$\begin{aligned}
\sigma^2(Y) &= \frac{(y_1 - y_2)^2}{n^2} + \frac{(y_1 - y_3)^2}{n^2} + \frac{(y_2 - y_3)^2}{n^2} \\
&= \frac{4}{n^2}\left[\frac{(y_1 - y_2)^2}{4}\right] + \frac{4}{n^2}\left[\frac{(y_1 - y_3)^2}{4}\right] + \frac{4}{n^2}\left[\frac{(y_2 - y_3)^2}{4}\right],
\end{aligned}$$

where the expected value of $\sigma^2(Y)$ is to be interpreted as $\sigma^2(\overline{T}_0)$. Note that each of the terms in square brackets has the form of Equation 2, which is the variance for a vector of two scores. It follows from the derivation of Equation 3 (see especially Equation 5) that

$$\begin{aligned}
\sigma^2(\overline{T}_0) &= \frac{4}{n^2}\left\{\left[\frac{n(n-1)}{2}\right]\frac{\sigma^2(T|S)/N}{2} + \sum\sum_{i<j}\frac{(\overline{T}_i - \overline{T}_j)^2}{4}\right\} \\
&= \left(\frac{n-1}{n}\right)\frac{\sigma^2(T|S)}{N} + \frac{1}{n^2}\sum\sum_{i<j}(\overline{T}_i - \overline{T}_j)^2, \qquad (8)
\end{aligned}$$

where $\overline{T}_i$ and $\overline{T}_j$ are the population mean scores for schools i and j, respectively. Now, the last term in Equation 8 has the form of a variance, as given by Equation 1. Thus,

$$\sigma^2(\overline{T}_0) = \left(\frac{n-1}{n}\right)\frac{\sigma^2(T|S)}{N} + \sigma^2(\overline{T}). \tag{9}$$

QED

References

Hill, R., & DePascale, C. (2002). *Determining the reliability of school scores.* Dover, NH: National Center for the Improvement of Educational Assessment, Inc.