RESEARCH
MEMORANDUM

July 2003
RM-03-02

An Interesting Problem in the Estimation of Scoring Reliability

Samuel A. Livingston

Research &
Development Division
Princeton, NJ 08541

An Interesting Problem in the Estimation of Scoring Reliability

Samuel A. Livingston

Educational Testing Service, Princeton, NJ

July 2003

Research Memoranda provide preliminary and limited dissemination of ETS research prior to publication. They are available without charge from:

Research Publications Office
Mail Stop 10-R
Educational Testing Service
Princeton, NJ 08541

Abstract

A performance assessment consisting of 10 separate exercises was scored with a randomized scoring procedure. All responses to each exercise were rated; then a randomly selected subset of the responses to each exercise received an independent second rating. Each second rating was averaged with the corresponding first rating before the scores were computed. This report presents a method for estimating the scoring reliability coefficient (inter-rater reliability) and the standard error of scoring of the resulting scores. The report concludes with a numerical example.

Key words: inter-rater reliability

Acknowledgements

Any opinions expressed in this paper are those of the author and do not necessarily reflect the position of Educational Testing Service or any of its clients.

The Problem

The problem discussed in this paper arises from an actual testing situation involving a performance assessment consisting of 10 separate exercises. (The solution generalizes easily to assessments of more or fewer than 10 exercises.)

Scores on the assessment are computed from numerical ratings of the candidate's responses to the 10 exercises. Each of a candidate's 10 responses receives a first rating. In addition, many of the responses to each exercise are randomly selected to receive a second rating. The second rating is made by a second rater, working independently and without knowledge of the first rating. The selection of the responses for double-scoring on any exercise is independent of the selection for double-scoring on any other exercise. If a response is single-scored, the candidate's score for the exercise is simply the numerical rating assigned by the rater. If a response is double-scored, the candidate's score for the exercise is the average of the ratings assigned by the two raters. The candidate's total score is a weighted sum of the scores on the 10 individual exercises. The responses to each exercise are rated by a different group of raters. Consequently, every rating contributing to a candidate's score is made by a different rater.

The problem is to estimate statistics that describe the reliability of the scoring process—a reliability coefficient and the associated standard error of measurement—from the data produced by an operational scoring of the assessment. These reliability statistics take into account only those sources of measurement error contributed by the scoring process. To emphasize this important point and to avoid confusion with statistics that include other sources of measurement error, this paper will refer to these statistics as the "scoring reliability" and the "standard error of scoring."

A Practical Approach

The scoring reliability coefficient is the correlation between the total scores that would result from two independent replications of the entire scoring process. *The random selection of the responses to each exercise that receive a second scoring is a part of the scoring process.* It contributes to the randomness of the process, just as the assignment of a candidate's response to a particular rater does. An estimate of the reliability of the scoring process must include both these sources of randomness. The approach taken in this paper explicitly incorporates the selection probabilities into the estimate.

For a test of 10 exercises, there are $2^{10} = 1{,}024$ possible combinations of single-scoring and double-scoring. Clearly, it is not practical to estimate the scoring reliability of each of the 1,024 possible combinations of single-scoring and double-scoring. Instead, the solution described in this paper takes the following approach:

1. For each exercise, estimate the scoring error variance separately for candidates whose responses were single-scored and those whose responses were double-scored.
2. Use the selection probabilities for single and double scoring to combine the separate estimates from Step 1, to obtain an estimate of the scoring error variance for the exercise in the group of all the candidates.
3. Use the scoring weights to combine the scoring error variance estimates for the individual exercises, to obtain an estimate of the scoring error variance of the total score. Use this result to estimate the scoring reliability coefficient and the standard error of scoring for the full test in the group of all candidates.

The Solution

If X_i is the random variable representing the candidate's score on exercise *i*, and w_i is the weight assigned to exercise *i*, then the random variable representing the candidate's total score is

$$X_{total} = \sum_i w_i X_i \ .$$

For the group of responses to exercise *i* that were single-scored, the variance of errors of scoring (VES) is simply the VES of the first ratings. However, there are no second ratings of these responses to use in estimating this quantity. It will be necessary to assume that some reliability statistic is the same in the single-scored group as in the double-scored group. In general, reliability coefficients tend to vary from group to group, but standard errors of measurement tend to be quite similar in different groups of candidates. Therefore, this solution will assume that the standard error of the first rating is the same in the single-scored group as in the double-scored group. Since the VES is simply the square of the standard error of scoring, this assumption is equivalent to assuming that

$$VES_{sngl}(X_{i,1st}) = VES_{dbl}(X_{i,1st}) ,$$

where $X_{i,1st}$ is the random variable representing the rating assigned by the first rater of a candidate's response to exercise *i*. (Similarly, $X_{i,2nd}$ is the random variable representing the rating assigned by the second rater. The subscripts *sngl* and *dbl* identify the single-scored and double-scored groups of responses.)

This quantity can be estimated from the variance of the first ratings and the correlation of first and second ratings in the double-scored group:

$$estdVES_{dbl}(X_{i,1st}) = \left[1 - r_{i,1st,2nd}\right]\mathrm{var}_{dbl}(X_{i,1st}) .$$

It seems reasonable to assume, in the absence of evidence to the contrary, that the VES of the first ratings is the same for the examinees whose responses were single-scored as for the examinees whose responses were double-scored. Therefore, the estimated VES of the first ratings in the double-scored group can also serve as the estimated VES of the first ratings in the single-scored group—unless it is larger than the observed variance of those ratings. Since the exercise scores of the single-scored group are simply the first ratings, the estimated VES of those exercise scores should be the smaller of the two quantities:

$$estdVES_{sngl}(X_i) = \min\left\{\left[1 - r_{i,1st,2nd}\right]\mathrm{var}_{dbl}(X_{i,1st}), \mathrm{var}_{sngl}(X_{i,1st})\right\} .$$

The exercise score for a double-scored response is the average of the two ratings. Because errors of scoring in the first and second ratings are independent, the VES of the exercise scores in the double-scored group is

$$\begin{aligned} VES_{dbl}(X_i) &= VES_{dbl}\left[(X_{i,1st} + X_{i,2nd})/2\right] \\ &= \frac{1}{4}\left[VES_{dbl}(X_{i,1st}) + VES_{dbl}(X_{i,2nd})\right]. \end{aligned}$$

The estimate of this quantity is

$$\begin{aligned} estdVES_{dbl}(X_i) &= \frac{1}{4}\left[\left(1 - r_{i,1st,2nd}\right)\mathrm{var}_{dbl}(X_{i,1st}) + \left(1 - r_{i,1st,2nd}\right)\mathrm{var}_{dbl}(X_{i,2nd})\right] \\ &= \frac{1}{4}\left(1 - r_{i,1st,2nd}\right)\left[\mathrm{var}_{dbl}(X_{i,1st}) + \mathrm{var}_{dbl}(X_{i,2nd})\right] . \end{aligned}$$

It seems reasonable to assume that the error of scoring in each candidate's exercise score is independent of the error of scoring in any other candidate's exercise score. This assumption enables us to combine the estimate of the VES of the exercise scores for the single-scored examinees with the estimate for the double-scored examinees, to get a VES estimate for the full set of responses to exercise i. To combine the VES estimates, we need to know the proportions of examinees whose responses to the exercise were single-scored and double-scored. These proportions will be denoted $p_{sngl}(X_i)$ and $p_{dbl}(X_i)$. Then the VES of the scores on exercise *i* in the full group of examinees is estimated by (the derivation of this formula appears in the appendix)

$$estdVES_{all}(X_i) = p_{sngl}(X_i) estdVES_{sngl}(X_i) + p_{dbl}(X_i) estdVES_{dbl}(X_i) .$$

Because errors of scoring on any two exercises are independent, the estimate of the VES for the entire test, in the full group of candidates is

$$estdVES_{all}(X_{total}) = \sum_i w_i^2 \left[estdVES_{all}(X_i) \right].$$

The estimated standard error of scoring for the full test is

$$estdSEM_{all}(X_{total}) = \sqrt{estdVES_{all}(X_{total})} \, ,$$

and the estimated scoring reliability coefficient is

$$estdrel_{all}(X_{total}) = 1 - \frac{estdVES_{all}(X_{total})}{\text{var}_{all}(X_{total})} \; .$$

An Example

The data for the following example comes from one of the assessments sponsored by the National Board for Professional Teaching Standards (2003). These assessments are intended as "a mechanism of certification that recognizes accomplished teachers." The assessment consists of three "portfolio" exercises, one "documented accomplishments" exercise, and six "assessment center” exercises. All 10 exercises are rated on the same rating scale, but the scoring weights differ. The data in Table 1 are from the Early and Middle Childhood/Physical Education assessment. The table shows how the estimation procedure is implemented, step by step.

Table 1

Reliability Estimation Procedure, Showing Intermediate Steps

						estd	estd	estd	estd			estd	wtd
			var_{sngl}	var_{dbl}	var_{dbl}	VES_{sngl}	VES_{sngl}	VES_{sngl}	VES_{dbl}			VES_{all}	VES_{all}
Ex	Wt	$r_{1st,2nd}$	$(X_{i,1st})$	$(X_{i,1st})$	$(X_{i,2nd})$	$(X_{i,1st})$	$(X_{i,2nd})$	(X_i)	(X_i)	p_{sngl}	p_{dbl}	(X_i)	(X_i)
1	16	0.65	1.00	1.00	0.66	0.36	0.23	0.36	0.15	0.67	0.33	0.29	73.16
2	16	0.71	0.72	0.71	0.78	0.21	0.22	0.21	0.11	0.66	0.34	0.17	44.06
3	16	0.55	0.54	0.50	0.58	0.22	0.26	0.22	0.12	0.67	0.33	0.19	48.01
4	12	0.73	0.56	0.58	0.55	0.16	0.15	0.16	0.08	0.68	0.33	0.13	18.82
5	6.67	0.68	0.76	0.77	0.84	0.24	0.26	0.24	0.13	0.78	0.22	0.22	9.65
6	6.67	0.48	0.78	0.61	0.72	0.32	0.37	0.32	0.17	0.75	0.25	0.28	12.48
7	6.67	0.43	0.63	0.55	0.60	0.31	0.34	0.31	0.16	0.72	0.28	0.27	12.05
8	6.67	0.31	0.64	0.47	0.57	0.32	0.39	0.32	0.18	0.73	0.27	0.28	12.65
9	6.67	0.56	0.42	0.44	0.37	0.19	0.16	0.19	0.09	0.74	0.26	0.16	7.28
10	6.67	0.44	0.47	0.42	0.32	0.24	0.18	0.24	0.10	0.72	0.28	0.20	8.88
Total										estd	VES		247.05
										estd	SES		15.72
											var		2100.10
										estd	rel		0.88

References

National Board for Professional Teaching Standards. (2003). *Assessments analysis report.* Princeton, NJ: Educational Testing Service.

Appendix A

In general, let Y and Z be discrete random variables with E(Y) = E(Z) = μ, and let

$X = Y$ with probability p_y

$X = Z$ with probability p_z

in such a way that the distributions of Y and Z are independent of the random choice between $X = Y$ and $X = Z$.

Then $\operatorname{var}(X) = p_y \operatorname{var}(Y) + p_z \operatorname{var}(Z)$.

Proof:

$$\operatorname{var}(X) = E(X-\mu)^2$$

$$= \sum_k (k-\mu)^2 \Pr\{X=k\}$$

$$= \sum_k (k-\mu)^2 \left[\Pr\{X=Y, Y=k\} + \Pr\{X=Z, Z=k\}\right]$$

$$= \sum_k (k-\mu)^2 \Pr\{X=Y, Y=k\} + \sum_k (k-\mu)^2 \Pr\{X=Z, Z=k\}$$

$$= \sum_k (k-\mu)^2 \Pr\{X=Y\}\Pr\{Y=k\} + \sum_k (k-\mu)^2 \Pr\{X=Z\}\Pr\{Z=k\}$$

(because the distributions of Y and Z are independent of Pr{$X = Y$} and Pr{$X = Z$})

$$= \Pr\{X=Y\}\sum_k (k-\mu)^2 \Pr\{Y=k\} + \Pr\{X=Z\}\sum_k (k-\mu)^2 \Pr\{Z=k\}$$

$$= p_y \operatorname{var}(Y) + p_z \operatorname{var}(Z) \ .$$