Research Report

Evaluating Targeted Double Scoring for the Performance Assessment for School Leaders Using Imputation and Decision Theory

ETS RR–23-01

Jing Miao
Sandip Sinharay
Chris Kelbaugh
Yi Cao
Wei Wang

December 2023

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

RESEARCH REPORT

Evaluating Targeted Double Scoring for the Performance Assessment for School Leaders Using Imputation and Decision Theory

Jing Miao, Sandip Sinharay, Chris Kelbaugh, Yi Cao, & Wei Wang

Educational Testing Service, Princeton, NJ, USA

In a targeted double-scoring procedure for performance assessments that are used for licensure and certification purposes, a subset of responses receives an independent second rating if the first rating falls into a preidentified critical score range (CSR) where an additional rating would lead to considerably more reliable pass-fail decisions. This study evaluates the CSRs using two approaches—one based on imputation of missing scores and the other based on statistical decision theory—using data from the Performance Assessment for School Leaders (PASL). Results from the evaluation indicate that the currently used CSRs are effective.

Keywords critical score range; decision consistency; linear regression; expected loss function

doi:10.1002/ets2.12363

Portfolio (or performance) assessments have been called for to provide direct evidence of teaching practice in licensure decisions, and studies have found that portfolio assessments contribute unique information in addition to traditional standardized tests (e.g., Wilson et al., 2014). However, such assessments are extremely costly to score because the tasks require extended responses that often include supporting artifacts (e.g., documents and video recordings). In practice, all responses typically receive at least two independent ratings (e.g., as recommended by Williamson et al., 2012), and some receive a third rating for adjudication if needed. It seems intuitive that when a candidate scores very high or very low on a task, having a second rating is not very likely to change the pass or fail outcome on the total test. However, for candidates with scores in the borderline area or critical score range (CSR), a second rating is more likely to change the outcome. The targeted double-score (TDS) procedure was developed based on this logic (Finkelman et al., 2008; Miao et al., 2021; Miao & Cao, 2019) and has been implemented for several testing programs, including the *PRAXIS*® Performance Assessment for Teachers (*PPAT*®) and the Performance Assessment for School Leaders (PASL).

The PASL includes three tasks, each with four steps (or parts). The tasks are scored at the step level using a 0- to 4-point scale (see Table 1), and the task score is the sum of the four step scores. The total test score (whose maximum value is 64) is the weighted sum of the task scores, with Task 3 weighted twice as much as the other two tasks.

Starting from Fall 2019, the PASL adopted the TDS procedure, in which task responses receive an independent second rating (i.e., operational Rating 2, and hereafter referred to as OR2) if their scores from the first rater (R1) fall in a prespecified CSR. The final score for the response is the average of the two ratings.[1] For those with R1 scores outside of the CSRs, about 10% are randomly sampled to receive an independent second rating (i.e., agreement sample Rating 2, and hereafter referred to as AR2). This random sample is referred to as the agreement sample. The OR2s and AR2s allow us to evaluate interrater agreement across the full score range (see Table 2).

The CSRs were originally defined based on a visual examination of the mean total test scores conditional on task scores. As an example, Figure 1 shows the mean test score conditional on Task 1 scores—the R1 scores were used to create the figure. The solid blue line represents the conditional means of R1 test score, with the band of ±1 *SD* marked by the two dotted blue lines. The green horizontal line represents the test cut score of 42. The two red dashed lines mark the CSR of 8 to 11, where an independent second rating is quite likely to change the pass-fail outcome for a test taker. This approach of defining CSRs based on a plot like Figure 1 is straightforward, but somewhat arbitrary, as one can choose a narrower CSR of 9 to 10 or a wider CSR of 8 to 11. It is, therefore, important to validate the CSRs with empirical analyses, considering both psychometric quality of the resulting scores and scoring cost.

Corresponding author: J. Miao, E-mail: jmiao@ets.org

Table 1 Structure of the Assessment

Task features	Task 1	Task 2	Task 3
Number of steps	4	4	4
Score range for each step	0–4	0–4	0–4
Maximum possible task score	16	16	16
Task weight	1	1	2

Table 2 Targeted Double-Scoring Procedure

R1 score	% of responses double scored	Purpose
Above the CSR	10	AR2, for evaluating inter-rater agreement
Within the CSR	100	OR2, averaged with R1 score for reporting
Below the CSR	10	AR2, for evaluating inter-rater agreement

Note. R1 = first rater; CSR = critical score range; AR2 = agreement sample Rating 2; OR2 = operational Rating 2.

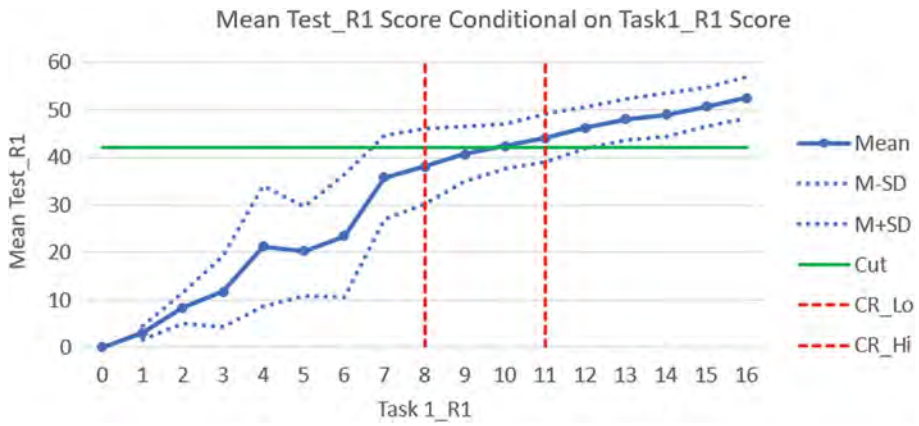

Figure 1 Mean Test_R1 score conditional on Task1_R1 score.

In this study, we used two different approaches to evaluate the CSRs—the missing data imputation approach and the decision theory-based approach. In the missing data imputation approach, we built a regression model to impute the missing scores resulting from the TDS approach and constructed synthetic double scores (SDSs) to be the criteria for evaluating the psychometric quality of different CSR options. The decision theory-based approach assumes certain losses for various decisions and applies statistical decision theory to observed data only to find optimum CSRs with the least loss. We sought to answer two questions:

1. Are the current CSRs effective (i.e., are we targeting double scoring at the right range)?
2. Are there better alternatives with lower cost but no deterioration in quality (i.e., do we need to adjust the current CSRs)?

Method: Evaluating the Critical Score Ranges Using Missing Data Imputation

Building the Regression Model

The data for this study were accumulated across 3 testing years from Fall 2018 to Spring 2021, with valid reported scores for 3,334 test takers. We first used all available data on double-scored task responses ($N = 971$, 993, and 1,230 for Tasks 1, 2, and 3, respectively) to build a regression model, which would generate a predicted Rater 2 (PR2) task score for all responses. The analysis was conducted separately for each task on data from the test takers whose task responses were scored by two human raters, R1 and R2 (either OR2 or AR2). Multiple linear regression models were built for each task, in which the R2 task score was the dependent variable and all 12 R1 step scores were predictors (or independent variables). Preliminary modeling analyses show that demographic variables such as gender and ethnicity did not improve model prediction and were therefore excluded from the model. To evaluate the prediction accuracy of the optimal regression model, root mean squared difference (RMSD) was computed as

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} \left(y_i - \hat{y}_i\right)^2}{N}}, \tag{1}$$

where N is the total number of examinees, y_i is the actual R2 score of examinee i on the task, and $\hat{y}_i$ is the PR2 score on this task. In addition, exact agreement (EA), exact plus adjacent agreement (E + AA), and quadratic weighted kappa (QWK) were also computed to examine the extent of agreement between the rounded PR2 task scores and the actual R2 task scores.

Because all the records in the data set were used as calibration data to build the model, to further validate the model, 10-fold cross-validation was conducted, where the model was built on 90% of the data and validated based on the remaining 10% of the data 10 times. In cross-validation, the general steps are as follows:

1. Divide the total double-scored sample into 10 equal groups in terms of sample size.
2. Use the first nine groups as the calibration sample to build the model and then predict the scores for the 10th group, which is used as the prediction sample. Compute RMSD.
3. Repeat Step 2 by using the ninth group, the eighth, and so on as the prediction samples and the remaining nine groups as the calibration samples.

The model was then used to predict the missing R2 task scores for examinees whose responses were single scored. Using the concept of prediction interval in linear regression (e.g., Draper & Smith, 1998, pp. 81–83), the final PR2 task scores were obtained as rounded (to the nearest integer) random draws from a normal distribution with the predicted values as the means and the standard errors of the predicted values of individual observations as the standard deviations. All the regression analyses were conducted using the R software.

The estimated regression coefficients in the final regression models for the three tasks are listed in Table 3, with coefficients printed in blue indicating statistically significant (at 5% level) predictors. The squared multiple coefficients for the three regressions were 0.16, 0.19, and 0.20, respectively.

Table 4 shows the RMSD, EA, E + AA, and QWK for each task based on averaged estimates from the 10 prediction samples. From Table 4, all the RMSD values are relatively high compared to the task scales (16 points). Although these linear models are the optimal ones we can obtain based on the available information, the RMSDs indicate that the models do not predict the second human scores very well. In addition, all the four agreement statistics are fairly low.

However, note that the values of the agreement measures, when computed from the full sample using the regression models described in Table 3, show a slightly better agreement, as shown in Table 5. For example, the values of QWK are between 0.18 and 0.19, which indicates a slight agreement (e.g., Landis & Koch, 1977) between the actual and predicted scores.

Table 6 provides the descriptive statistics of PR2 scores as compared to R1 scores. The means of PR2 scores are slightly lower than those of R1 scores (by −.22, −.04, and −.02, respectively), and the standard deviations appear to be very similar.

Table 3 Regression Models With Coefficients for Predicted Rater 2 Scores

			Task 1				Task 2				Task 3			
Predicted score	*N*	α	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
T1PR2	971	5.46	.29	.06	.26	.35	.17	.20	.17	.00	.31	−.07	.14	.16
T2PR2	993	4.45	.33	−.15	.38	.25	.16	.38	.20	.34	.18	.20	.04	.09
T3PR2	1,230	3.81	.27	.14	.23	.12	.14	.26	.11	.17	.47	.06	.23	.38

Note. α represents the regression intercept. T1PR2 = Task 1 predicted Rater 2; T2PR2 = Task 2 predicted Rater 2; T3PR2 = Task 3 predicted Rater 2; S1 = Step 1; S2 = Step 2; S3 = Step 3; S4 = Step 4. Coefficients printed in blue indicate statistically significant (at 5% level) predictors.

Table 4 Measures of Model Prediction Accuracy From the Validation Samples

Task	RMSD	EA	EA + AA	QWK
Task 1	2.56	0.13	0.45	0.16
Task 2	2.64	0.13	0.42	0.09
Task 3	2.32	0.12	0.43	0.15

Note. RMSD = root mean squared difference; EA = exact agreement; E + AA = exact plus adjacent agreement; QWK = quadratic weighted kappa.

Table 5 Measures of Model Prediction Accuracy From the Full Sample

Task	RMSD	EA	EA + AA	QWK
Task 1	2.42	0.17	0.46	0.19
Task 2	2.51	0.16	0.46	0.18
Task 3	2.38	0.17	0.48	0.19

Note. RMSD = root mean squared difference; EA = exact agreement; E + AA = exact plus adjacent agreement; QWK = quadratic weighted kappa.

Table 6 Descriptive Statistics for Rater 1 (R1) Score and Predicted Rater 2 (PR2) Score

Score	*N*	Mean	*SD*	Min	Max
Task 1 R1	3,334	10.97	2.79	0	16
Task 1 PR2	3,334	10.75	2.82	0	16
Task 2 R1	3,334	10.86	2.99	0	16
Task 2 PR2	3,334	10.82	2.98	0	16
Task 3 R1	3,334	10.59	3.03	0	16
Task 3 PR2	3,334	10.57	2.98	0	16

Compiling the Synthetic Double-Score Data

We then compiled the SDS data for analyses to evaluate the efficacy of the CSRs. We computed the synthetic Rater 2 (SR2) task score as follows:

- SR2 equals to the OR2 score if available;
- SR2 equals to the AR2 if available;
- SR2 equals to the PR2 score if neither OR2 nor AR2 is available.

The SDS task score is the average[2] of R1 and SR2, and the SDS test score is the weighted sum of the SDS task scores. We used SDS scores as the criteria for comparisons.

Evaluating the Critical Score Ranges

Using the graphic tool introduced earlier, we first identified five task-level CSRs[3] for evaluation. We applied these CSRs to calculate test scores and pass-fail outcomes under each condition specified in Table 7. To determine the optimal option, we compared the results on alpha reliability, pass rate, agreement of classification, and scoring scope, with the SDS condition results as the criteria. The scoring scope is the total number of scores produced for each condition as a percentage of the total number of scores produced in the 7 to 10 baseline condition.[4]

The operational procedure uses the 7 to 10 CSRs for all three tasks; a small number of responses may receive third-rater adjudication when the difference between R1 and R2 exceeds the predefined allowable difference.[5] In this study, we only considered R1 and R2 scores without third-rater adjudication; therefore there are small differences between the "operational" and the "7 to 10" condition. Results in Table 7 show that the current 7 to 10 CSR is the optimal option among the group of five, if we are making double-score decisions on R1 task score. The more costly options do not improve the psychometric quality; the narrower range of 9 to 10 has lower cost but shows deterioration in quality.

Table 7 Targeted Double-Scoring Conditions With Task-Level[6] Critical Score Ranges

Scoring condition	Alpha reliability	Pass rate (%)	% agree with SDS criteria	Scoring scope[7] (%)
SDS (criteria)	0.84	80		
Operational	0.83	85	93.6	
7 to 10	0.83	83	93.8	100
7 to 11	0.83	78	92.9	111
8 to 11	0.83	78	92.9	110
9 to 11	0.82	74	90.3	106
9 to 10	0.83	75	89.0	95

Note. SDS = synthetic double scores. Green shading indicates that 7 to 10 is the optimal option among the group of five. Green indicates desirable values, while red indicates the opposite.

Table 8 Targeted Double-Scoring Conditions With Test-Level Critical Score Ranges

Scoring condition	Alpha reliability	Pass rate (%)	% agree with SDS criteria	Scoring scope (%)
SDS (criteria)	0.84	80		
Operational	0.83	85	93.6	
37 to 41	0.81	83	88.0	94
39 to 42	0.81	81	96.2	96
37 to 42	0.82	80	97.3	98
37 to 44	0.82	80	99.1	107
37 to 45	0.83	80	99.6	111

Note. SDS = synthetic double scores. Green shading indicates that 39 to 42 is the optimal option among the group of five. Green indicates desirable values, while red indicates the opposite.

Table 9 Targeted Double-Scoring Conditions with Task-Level and Test-Level Critical Score Ranges

Scoring condition	Alpha reliability	Pass rate (%)	% agree with SDS criteria	Scoring scope (%)
SDS (criteria)	0.84	80		
Operational	0.83	85	93.6	
(8 to 11) × (37 to 42)	0.83	82	94.5	93
(7 to 10) × (37 to 45)	0.82	83	93.7	95
(8 to 11) × (37 to 44)	0.82	82	94.5	98
(7 to 11) × (37 to 44)	0.82	82	95.0	98
(7 to 11) × (37 to 45)	0.82	82	95.1	101

Note. SDS = synthetic double scores. Green shading indicates that (8 to 11) × (37 to 42) is considered the optimal option among the group of five.

In search of more cost-efficient alternatives (i.e., to answer the second research question), we also considered CSRs using *test* scores. One approach is to use R1 test score CSR; if a candidate's weighted sum of three R1 task scores falls in a predefined CSR, all tasks will be double scored. Table 8 provides results from five test-level CSRs right below or around the cut score of 42. Balancing considerations of quality and cost, we considered 39 to 42 to be the optimal from this group of five, if we are making double-score decisions on R1 test score.

Another approach is to use dual-level CSRs considering both the task-level and the test-level scores; i.e., a task response will get a second score if both the R1 task score and R1 test score fall in the respective CSRs. Considering results from task-level and test-level analysis, we analyzed different dual-level combinations. Table 9 provides results from five conditions with better performances. The optimal one in this group of five is the combination of task-level CSR of 8 to 11 and test-level CSR of 37 to 42. It has the lowest scoring cost among the five options, which are of similar psychometric qualities.

Altogether, we compared a total of 15 sets of CSRs (Tables 7–9) using three different approaches and identified the optimum CSR for each. Table 10 provides the summary statistics of task and test scores of the three optimum scoring conditions. Single-score (R1) statistics are provided for reference. Overall, the different conditions produce similar score distributions; TDS results tend to have slightly higher means than SDS and single-scored (R1) conditions; single-scored (R1) scores have largest standard deviations; and the TDS scores have *SD*s between SDS and single-scored conditions.

Table 10 Descriptive Statistics of Task and Test Scores Under Different Scoring Conditions

Task	Scoring conditions	*N*	Mean	*SD*	Min	Max
Task 1	SDS (criteria)	3,334	10.86	2.58	0	16
	TDS operational	3,334	11.11	2.73	0	16
	TDS Condition 1: Task (7 to 10)	3,334	11.11	2.73	0	16
	TDS Condition 2: Test (39 to 42)	3,334	10.97	2.76	0	16
	TDS Condition 3: (8 to 11) × (37 to 42)	3,334	11.02	2.77	0	16
	Single score (R1)	3,334	10.97	2.79	0	16
Task 2	SDS (criteria)	3,334	10.85	2.80	0	16
	TDS operational	3,334	11.01	2.93	0	16
	TDS Condition 1: Task (7 to 10)	3,334	11.01	2.93	0	16
	TDS Condition 2: Test (39 to 42)	3,334	10.88	2.97	0	16
	TDS Condition 3: (8 to 11) × (37 to 42)	3,334	10.92	2.97	0	16
	Single score (R1)	3,334	10.86	2.99	0	16
Task 3	SDS (criteria)	3,334	10.59	2.82	0	16
	TDS operational	3,334	10.78	2.97	0	16
	TDS Condition 1: Task (7 to 10)	3,334	10.78	2.97	0	16
	TDS Condition 2: Test (39 to 42)	3,334	10.68	2.99	0	16
	TDS Condition 3: (8 to 11) × (37 to 42)	3,334	10.72	2.98	0	16
	Single score (R1)	3,334	10.59	3.03	0	16
Test	SDS (criteria)	3,334	43.13	10.26	0	60
	TDS operational	3,334	43.75	10.67	0	62
	TDS Condition 1: Task (7 to 10)	3,334	43.69	10.68	0	62
	TDS Condition 2: Test (39 to 42)	3,334	43.25	10.69	0	62
	TDS Condition 3: (8 to 11) × (37 to 42)	3,334	43.38	10.66	0	62
	Single score (R1)	3,334	43.01	10.70	0	62

Note. SDS = synthetic double scores; TDS = targeted double scores; R1 = Rater 1.

Table 11 Reliability Estimates and Classification Agreement Under Different Scoring Conditions

Scoring condition	Alpha	Decision accuracy	Decision consistency	Pass rate (%)	% agree with SDS	Scoring scope (%)
SDS	0.84	0.90	0.86	80		
TDS Operational	0.83	0.86	0.85	85	93.6	
TDS Condition 1	0.83	0.86	0.85	83	93.8	100
TDS Condition 2	0.81	0.88	0.84	81	96.2	96
TDS Condition 3	0.83	0.86	0.84	82	94.5	93

Note. SDS = synthetic double scores; TDS = targeted double scores. Green indicates desirable values.

Table 11 provides the reliability estimates for the three conditions, which have similar yet slightly lower values than the SDS estimates. Alpha reliability is a measure of the internal consistency of the test and is related to test length. In our data, each task included four steps scored by the same rater, and therefore the steps cannot be treated as independent test items. Our estimates are based on the test length of three items and provide a lower limit of alpha reliability. Reliability of classification measures are calculated when a cut score is used for pass or fail decisions (Livingston & Lewis, 1995). Decision accuracy refers to the extent to which the classifications of test takers based on their scores on the test form agree with the classifications made using perfectly reliable test scores. Decision consistency refers to the agreement between these classifications based on two nonoverlapping, equally difficult forms of the test. By definition, decision consistency values are always lower than the corresponding decision accuracy values.

Table 11 also provides pass rates and classification agreement rates, which are more tangible measures to nontechnical stakeholders. The pass rate ranges from 80% for the SDS condition to 85% for the operational condition, whereas the three TDS conditions have similar pass rates in between. The classification agreement rate ranges from 93.8% for Condition 1 and 96.2% for Condition 2, all above 93.6% agreement rate for the operational condition. The three TDS conditions appear to be similar in quality, but Condition 3 can be considered the optimum for its lowest cost (see Table A2 in the appendix for calculation details).

Method: Evaluating the Critical Score Ranges Using Decision Theory Analysis

The Motivation Behind the Approach

We decided to also apply statistical decision theory (e.g., Rudner, 2009) to choose optimal CSRs because this approach only utilizes available data and does not involve the use of imputed data. In applications of statistical decision theory (henceforth referred to as *decision theory*), an investigator chooses one among a set of possible actions or decisions. A major tool in decision theory is a loss function that quantifies the gain or loss associated with each possible action. An application of decision theory also involves an unknown element that is typically expressed as a probability distribution. The investigator computes the average (or expected) loss averaged over the probability distribution for each action and chooses the action that leads to the minimum average loss among all possible actions.

The Expected Loss Function

In the context of this paper, an action corresponds to the choice of a specific CSR for a task, and the unknown elements are the R1 and R2 scores on that task and the other tasks of the PASL. If the R1 score of an examinee lies within the CSR (that leads to the double scoring for the examinee on the task), the resulting loss is assumed to be equal to the cost of an extra rating on the task. Let us denote this cost as c. For each R1 score (on the task of interest) that lies outside the CSR, no double scoring is performed, but that event is assumed to lead to a loss of

- L_P if the examinee passes the PASL based on Rater 1 task score (that is, with single scoring) and would have failed the test if the task were double scored;
- L_F if the examinee fails the PASL based on Rater 1 task score (with single scoring) and would have passed the test if the task were double scored;
- 0 if the examinee's pass-fail status is the same irrespective of whether one or two ratings are used on the task.

The quantity L_P primarily quantifies the loss that borderline examinees who incorrectly passed the PASL would cause by performing poorly at their professions. The quantity L_F represents the loss corresponding to the potentially unfair failing and the resulting loss of job income of a borderline examinee. This formulation of the problem capitalizes on the fact that double scoring leads to an improvement in the quality of the scores compared to single scoring, which has been noted by, for example, Williamson et al. (2012). The expected loss corresponding to a CSR is then given by

$$\begin{aligned} \mathrm{EL}_{\mathrm{CSR}} = \mathrm{c}\mathrm{P}(\mathrm{CSR}) + \big[&\mathrm{L}_{\mathrm{P}}\ \mathrm{P}(\text{Pass with 1 rating and fail with 2 ratings \& Outside CSR}) \\ &+ \mathrm{L}_{\mathrm{F}}\ \mathrm{P}(\text{Fail with 1 rating and pass with 2 ratings \& Outside CSR})\big]\ [1 - \mathrm{P}(\mathrm{CSR})], \end{aligned} \tag{2}$$

where P(CSR) denotes the probability that the Rater 1 task score of a randomly chosen PASL examinee falls within the CSR and where P(Pass with 1 rating and fail with 2 ratings & Outside CSR) is the probability of a random examinee passing the PASL based on the R1 score and failing the test if the task were double scored given that the examinee's R1 score on the task lies outside the CSR.

Equation 2 can be used to estimate the expected loss of each CSR if the probabilities in the equation can be accurately estimated. If a representative sample of double-scored responses for the task of interest is available, then the probabilities in Equation 2 can be estimated by the corresponding sample proportions; the sample of double-scored examinees that were used in the imputation-based approach (described in the previous subsection) was used as the representative sample to estimate the probabilities in the above equation.[8] The CSRs that led to the smallest estimated expected losses were chosen as the optimum CSRs. The approach of choosing the CSR by minimizing the expected loss constitutes an optimal approach in the sense that (a) such an approach has desirable theoretical properties and (b) situations can be constructed in which the follower of other approaches will be assured of inferior results. More details about the approach can be found in Sinharay et al. (2022).

Results From Application of the Decision Theory-Based Approach to Our Data

To estimate the expected loss using the above equation, one needs appropriate values of c, L_P, and L_F—these values are problem specific. Conversations with the administrators of the test made it clear that one task rating costs about

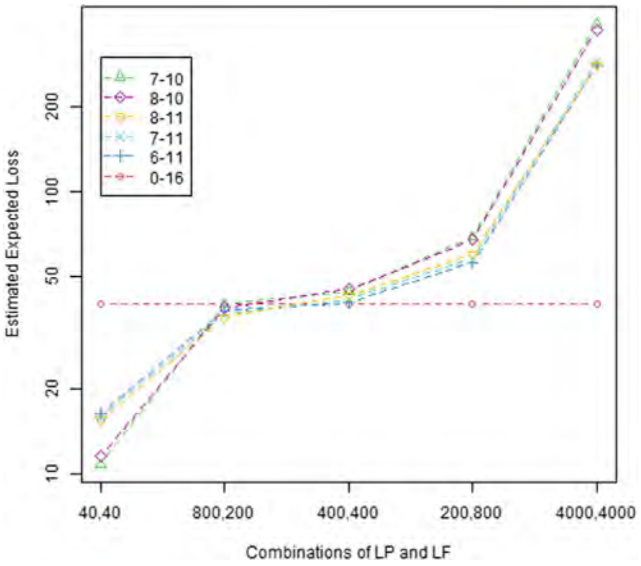

Figure 2 Expected losses for various critical score ranges and various combinations of L_P and L_F.

$40 on average after including all expenses. Therefore, c was set equal to 40. The following five combinations of L_P and L_F (each of which represents cost in dollars) were used in the computations of the estimated expected losses: {40,40}, {400,400}, {800,200}, {200,800}, and {4,000,4,000}. The first and fifth combinations were chosen to represent two extremes, representing the cost of an incorrect decision (arising out of single scoring) to be very low and very high, respectively. The other three combinations represent more moderate conditions. Although $L_P = L_F$ in three of the five combinations, the combinations {800,200} and {200,800} were chosen to represent unequal losses from the two types of incorrect decisions. The values, such as 200, 400, and 800 and so on, were chosen after a discussion with those familiar with the testing program and roughly represent realistic losses caused by various decisions.

Figure 2 shows the estimated expected losses along the vertical axis for Task 1 for six CSRs that include the operational CSR for the task (7 to 10), one extreme CSR (the set 0–16 representing double scoring of all responses), and four other CSRs that were found to lead to small estimated expected losses on average in a preliminary investigation. Each line corresponds to a CSR and connects five points that represent the estimated expected losses for five combinations of L_P and L_F that are represented along the horizontal axis for that CSR. The six CSRs are denoted using different symbols, as denoted in the legend of the figure. The figure and similar figures for the other two tasks indicate that 6 to 11, 7 to 11, and 8 to 11 led to the minimum estimated expected loss for all tasks.

We also performed another set of decision theory-based analysis using CSRs based on both the task scores and the total R1 score[9]—this analysis revealed that the minimum estimated expected loss for a task was achieved for the CSR that is 8 to 11 for the task and 37 to 42 for the total R1 score (the results were very close but slightly worse for the CSR that is 7 to 11 for the task and 37 to 42 for the total R1 score). In addition, the minimum estimated expected losses from this analysis were on average 10–20% smaller than that for CSRs based only on task scores (and not on total scores).

Therefore, the decision theory-based analysis indicates that revising the double-scoring policy to double score a response when the Task R1 score is between 8 and 11 (or 7 to 11) and the total R1 score is between 37 to 42 is optimal from a decision-theoretic point of view. This is consistent with the results from the imputation-based approach.

Summary

In this study we applied two approaches to evaluate the CSRs currently used for PASL and also to search for more cost-efficient alternatives. One approach imputed missing data to create an SDS as the criterion for evaluation; the other used only observed data but made some assumptions on the losses associated with different conditions. The two approaches arrived at converging results.

The current task level CSRs of 7 to 10 were shown to be adequate in terms of psychometric quality. The decision theory-based analysis results indicate that 6 to 11, 7 to 11, and 8 to 11 led to the lower estimated expected loss than

7 to 10; however, the scoring cost corresponding to each of these three CSRs was 10–11% higher than the current procedure.

The dual-level CSRs (i.e., task level 8 to 11 in combination with R1 test score 37 to 42) have the potential to reduce the scoring cost by 7%; however, the implementation would involve substantial reconfigurations to the current scoring system and work processes. It needs to be determined whether the cost of implementing such change can be offset by the potential savings from reduced scoring scope.

TDS is a cost-effective procedure for performance assessment where automated scoring is not yet feasible. This study provides a framework for establishing and evaluating the CSRs where a full double-scored criterion is not available. This approach can be adopted when designing new assessments with constructed responses that cannot be easily scored automatically. It is also possible to apply our methods and analysis to data from other tests and simulated data, including tests that have items with various formats and various types.

Our study has several limitations. First, although we considered score reliability and did not consider validity in this paper, it is possible to compare CSRs with respect to predictive validity if an external criterion is available. Second, although we used linear regression to impute missing data, other advanced methods for missing-score imputation (e.g., those from Sinharay, 2021) can be applied in future research.

Notes

[1] For a small number of cases, a third rating may be needed for adjudication if the first two ratings are discrepant.

[2] Third rater adjudication is used in a very small number of cases and not considered in this study.

[3] CSRs can be different for each task, but they happen to be the same for all three tasks in this case.

[4] Detailed calculation is provided in Table A1.

[5] The allowable difference is 4 points for task 1 and task 2, and 3 points for double-weighted task 3.

[6] Task level CSRs can be different for each task, but they happen to be the same for the data in this study.

[7] See Table A1 for more information on scoring scope calculation.

[8] While estimating the probabilities, weighting was used to account for the fact that all responses with Rater 1 score in the CSR are double scored. butle only 10% of the responses with R1 score outside the CSR are double scored.

[9] An analysis was also performed after defining the CSRs based on only the task scores—the expected losses were not smaller than those in Figure 2—so CSRs based only on task scores were not considered any more.

References

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781118625590

Finkelman, M., Darby, M., & Nering, M. (2008). A two-stage scoring method to enhance accuracy of performance level classification. *Educational and Psychological Measurement*, *69*, 5–17. https://doi.org/10.1177/0013164408322025

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179–197.

Miao, J., & Cao, Y. (2019, April 5–9). *Development and evaluation of a partial double scoring procedure for preservice teacher portfolio assessment* [Paper presentation]. American Educational Research Association Annual Meeting, Toronto, Canada. https://doi.org/10.3102/1432878

Miao, J., Wang, W., Sinharay, S., Kelbaugh, C., & Cao, Y. (2021, May 18–June 11). *Evaluating targeted double scoring for performance assessment using simulated* data [Paper presentation]. National Council on Measurement in Education Annual Meeting, virtual.

Rudner, L. M. (2009). Scoring and classifying examinees using measurement decision theory. *Practical Assessment, Research, and Evaluation*, *14*, 1–12.

Sinharay, S. (2021). Score reporting for examinees with incomplete data on large-scale educational assessments. *Educational Measurement: Issues and Practice*, *40*(1), 79–91. https://doi.org/10.1111/emip.12396

Sinharay, S., Johnson, M., Wang, W., & Miao, J. (2022). Targeted double scoring for performance assessment using a decision-theoretic approach: A case study. *Applied Psychological Measurement*. https://doi.org/10.1177/01466216221129271

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practices*, *31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Wilson, M., Hallam, P. J., Pecheone, R. L., & Moss, P. A. (2014). Evaluating the validity of portfolio assessment for licensure decisions. *Education Policy Analysis Archives*, *22*(6). https://doi.org/10.14507/epaa.v22n6.2014

Appendix

Table A1 Estimated Scoring Scope for Three Sets of Task-Level Critical Score Ranges

CSR	Score level	*N* of R1	*N* of OR2	*N* of AR2	Total scoring	% of baseline
7 to 10	Task 1	3,210	643	257		
	Task 2	3,194	672	252		
	Task 3	3,186	910	228		
	Test	9,590	2,225	737	12,552	100
8 to 11	Task 1	3,210	1,098	211		
	Task 2	3,194	1,148	205		
	Task 3	3,186	1,359	183		
	Test	9,590	3,605	599	13,794	110
7 to 11	Task 1	3,210	1,129	208		
	Task 2	3,194	1,172	202		
	Task 3	3,186	1,393	179		
	Test	9,590	3,694	589	13,873	111

Note. CSR = critical score range; R1 = Rater 1; OR2 = optimal Rating 2; AR2 = agreement sample Rating 2. Red indicates undesirable values.

Table A2 Estimated Scoring Scope for Specified Targeted Double-Score Conditions

Scoring condition	Score level	*N* of R1	Task-level CSR	Test-level CSR	*N* of OR2	*N* of AR2	Total scoring	% of baseline
#1	Task 1	3,210	7 to 10	NA	643	257	4,110	NA
	Task 2	3,194	7 to 10		672	252	4,118	
	Task 3	3,186	7 to 10		910	228	4,324	
	Test	9,590			2,225	737	12,552	100
#2	Task 1	3,210	NA	39 to 42	541	267	4,121	NA
	Task 2	3,194	NA		541	265	4,103	
	Task 3	3,186	NA		541	265	4,094	
	Test	9,590			1,623	797	12,010	96
#3	Task 1	3,210	8 to 11	37 to 42	349	286	3,845	NA
	Task 2	3,194	8 to 11		371	282	3,847	
	Task 3	3,186	8 to 11		595	259	4,040	
	Test	9,590			1,315	827	11,732	93

Note. R1 = Rater 1; CSR = critical score range; OR2 = optimal Rating 2; AR2 = agreement sample Rating 2.

Suggested citation:

Miao, J., Sinharay, S., Kelbaugh, C., Cao, Y., & Wang, W. (2023). *Evaluating targeted double scoring for the Performance Assessment for School Leaders using imputation and decision theory* (Research Report No. RR-23-01). ETS. https://doi.org/10.1002/ets2.12363

Action Editor: Gautam Puhan

Reviewers: Jodi Casabianca and Michael Walker