

## A Comparison of Distributed and Regional Scoring

Leslie Keng, Laurie L. Davis and Shelley Ragland

Distributed scoring is a model of performance scoring in which readers receive training and conduct scoring remotely (e.g., from home) rather than in a regionally-located performance scoring center. Distributed scoring provides access to a wider pool of readers than those that could be included through regional scoring alone, thereby allowing for a larger number of readers to be recruited and permitting greater selectivity in reader recruitment. This has the potential for increased efficiency in training time for readers and could facilitate shorter turnaround times in performance scoring, which would, in turn, shorten the time between test administration and the reporting of test scores.

To evaluate the feasibility of implementing distributed scoring in the Texas Assessment Program, Pearson designed and conducted a research study that directly compared the regional and distributed scoring models using Texas training materials and Texas student responses. The study examined the practical impacts to the scoring process as well as to student essay scores and performance classifications under scoring conditions that sought to replicate the live scoring environment. The models were compared within the framework of the Texas Assessment Program in order to account for testing circumstances and processes that are unique to Texas.

The study was conducted on two sets of student responses to the essay item on the Texas Assessment of Knowledge and Skills (TAKS) exit-level English Language Arts (ELA) test administered in March 2009. The first set included a sample of 3,999 handwritten essay responses by students who took the primary administration of the ELA test; the sample of **primary testers** was selected to be representative of statewide population. The second set included *all* 1,291 essay responses from **online retesters** during the March administration.

*"To evaluate the feasibility of implementing distributed scoring in the Texas Assessment Program, Pearson designed and conducted a research study..."*

The two scoring conditions compared were **Distributed Scoring** in which readers conducted scoring remotely and **Regional Scoring** in which readers conducted scoring in a central performance-scoring center. A total of 157 readers participated in this study. The readers were experienced Texas readers who had scored regionally out of the Austin Performance Scoring Center. The readers were assigned to one of the two scoring conditions based on demographic characteristics (such as amount of scoring experience, level of education, type of degree, gender,

ethnicity and age group) to mitigate any potential influence of reader demographic characteristics on the study outcome. Readers assigned to the Distributed Scoring condition received asynchronous online training; while readers assigned to the Regional Scoring condition received stand-up training simultaneously at the performance scoring center.

The scoring process implemented for each scoring condition in the study mirrored that of the operational TAKS administration. Two initial readers (either both distributed or both regional) were randomly assigned to score every student's essay response. If the scores assigned by the two readers agreed, then the agreed upon score was assigned as the final score. If the scores for the two readers did not agree, a third resolution reader was assigned to score the essay. If the resolution reader assigned a score that agreed with one of the first two readers, then that score would be assigned as the final score. Otherwise, the essay was given to a senior scoring director, who assigned the final score for the essay. The scoring process for both conditions commenced at the same time and took place concurrently during a one-week period. This was done so that issues and challenges that occur during the scoring process could be captured and compared in real time. The entire scoring process was replicated for a second week to evaluate any effects due to the particular set of readers used in a given week. The readers in each condition were assigned to score in either the first or second week.

The two conditions were compared on traditional reader performance metrics, including agreement rates on practice and qualifying sets, rates of scoring, inter-rater agreements and validity agreements. They were also compared on the distribution of final essay scores after each essay had gone through the entire scoring process. Finally, impact analysis was conducted for each scoring condition by combining each student's final essay score in the study with the student's operational performance on the multiple-choice and open-ended sections to obtain a final ELA test score. The students were then classified into the TAKS performance levels (Did Not Meet, Met the Standard, Commended Performance) according to their final test score in each condition.

*"The study produced remarkably similar results for the two scoring conditions."*

The study produced remarkably similar results for the two scoring conditions. The results were also similar across the two study replications (weeks), implying that the readers used in each week did not have a significant impact on the outcome. As such, the study results were combined across the two weeks.

The distributed and regional readers scored a similar number of papers per hour and had similar agreement rates, inter-rater reliability and validity

agreement rates. This implies that there are no significant differences between distributed and regional scoring in the quantity and quality of scoring by the readers.

The distributions of the final scores given in each scoring condition for the primary testers and retesters show that perfect agreement was found across the distribution of final essay scores for primary testers; and near perfect agreement was found across distribution of final essay scores for retesters. These results imply that the final scores obtained through the Texas scoring process are not substantially different for the distributed and regional scoring models.

When comparing the impact analysis of the two scoring condition for the primary testers and retesters, a 95-98 percent consistency of student classification was found at the total test level between regional and distributed scoring. This means that the vast majority of students received the same classification under the two scoring conditions. Thus, the impact on Texas students in terms of their final performance level classification is strikingly similar under the distributed and regional scoring models.

There are some limitations to the study. First, the study was conducted on ELA essay papers for test-takers in TAKS. The extent to which the sample was representative of the general test-taking population in Texas affects the degree to which the results could be generalized to other testing programs in Texas. Also, the study duration was

short relative to operational scoring projects. Consequently, some reader performance metrics (e.g., inter-rater reliability, rate of scoring, etc.) may be artificially suppressed. Finally, the scoring process implemented in the study was similar to operational context, but did not perfectly replicate it (e.g., did not include specialist or analytic scoring). This may have had an effect on the study results, particularly for the retesters. Even with these limitations, the study results demonstrate that the use of distributed scoring can yield highly similar performance scoring results to the current regional scoring model in terms of reader performance, final score distribution and impact on student classification in Texas.

These results combined with its potential benefits provide compelling evidence supporting distributed scoring as a viable model in meeting the performance scoring requirements within the Texas Assessment Program in the years to come.