

APPENDIX 4

TEXAS OBSERVATION PROTOCOLS (TOP) WRITING STUDY, SPRING/SUMMER 2005

Texas Observation Protocols (TOP) Writing Study

Spring/Summer 2005

Introduction/Background

The Texas Education Agency's plans for evaluating the validity and reliability of the TOP ratings include periodic audits of the ratings for the domain of writing. Such audits will help the state evaluate rating validity and reliability as well as the impact of enhancements made in the training and rating process. The fact that TOP is a classroom observational assessment makes studies of interrater reliability in listening, speaking, and reading impractical to operationalize. Audits of rating validity and reliability will be conducted via the writing domain because the TOP assessment includes collections of student writing that can be rerated without the need for the state to conduct a large number of classroom observations over time. The first study of TOP writing ratings was conducted after the spring 2005 TOP administration.

In the month prior to the spring TOP administration window, TOP raters received holistic rating training through a statewide training-of-trainers model. A one-day training session was held in February 2005 for lead trainers from school districts and education service centers. Using materials provided by the state, the lead trainers were required to train other rater trainers or the raters themselves. The training materials included an in-depth module on the rating rubrics; annotated examples of student performance in the domains of listening, speaking, and writing; and authentic student writing samples for raters to use to practice applying the holistic rubrics.

In the administration of TOP, raters were directed to holistically rate the writing proficiency level of students in Grades 2–12 based on classroom observations of the students over time. Raters were also instructed to assemble a collection of 3–5 writing samples per student. These writing collections were to reflect the overall writing abilities of their students and support their broader observational ratings in the event that their writing collections were selected for a state audit.

The TOP writing study included three main activities:

- 1) Determining the agreement rates between teacher and state ratings assigned to the audited writing collections
- 2) Summarizing the results of a survey administered to the teachers who rated the audited writing collections
- 3) Evaluating the adequacy of the writing collections for use in judging the overall writing proficiency of English language learners

Method

In May 2005 following the TOP administration, PEM staff collected a stratified random sample of the teacher-rated student writing collections. Writing collections from thirty students were collected per grade band (2-3, 4-5, 6-8, and 9-12) and teacher-assigned proficiency level (beginning, intermediate, advanced, and advanced high.) This resulted in a sample size of $30 * 4 * 4 = 480$ students. A total sample size of 400 was desired, so over-sampling was conducted with the goal of obtaining 25 writing collections per grade band and proficiency level. A comparison of the sample and the full population is shown in Table 1. Input on this sampling design was received from Texas Technical Advisory Committee (TTAC) members before it was implemented.

After the writing samples were received, a team of TEA and Pearson personnel involved in the TELPAS program met and independently re-rated the writing samples without knowledge of the teacher assigned ratings. This rating team had extensive experience developing the TOP rating rubrics and training materials. The ratings of the individuals on this team were used as a criterion to compare against the sampled teacher ratings.

Results

Of the 480 student writing collections requested, the school districts submitted 461 writing collections. Of the 461 received collections, 114 (25%) were classified as “unscorable” by the rating team. Collections were deemed unscorable if they could not be used to judge the overall writing ability of a student because of an insufficient number, variety, or type of writing samples. A total of 347 writing collections were able to be re-rated by the state.

Table 2 shows the distribution of ratings by teachers and by the state rating team. The number of collections received is relatively uniform across the proficiency levels, suggesting that the 19 collections not submitted by the school districts were random. Although the sample was designed to obtain equal numbers of collections across grade bands and teacher-assigned proficiency level, the majority of the submitted collections were found to be at the intermediate or advanced levels by the state. According to the state ratings, the sample included far fewer beginning collections and slightly fewer advanced high collections than was expected given the teacher ratings.

Table 3 displays the exact agreement rates between the teacher and the state rater by grade band. Overall, the teacher rating agreed with the state rating 42% of the time. There was little variation in the agreement rate across grade bands, suggesting that grade level had little or no impact on the rating accuracy.

Table 4 presents the exact agreement rates between the state and teacher ratings. The state ratings agreed with the teacher ratings 42% of the time. The agreement rate was lowest (38%) for the writing collections teachers rated as beginning. The agreement rates for the collections teachers rated as intermediate, advanced, and advanced high were 43%, 43%, and 44%, respectively.

Table 5 presents a cross-tabulation of teacher rating by state rating. In addition to the 42% of the collections which received the same rating by the teacher and the state, this analysis

indicates that 25.95% of the writing collections were rated higher by the teacher than the state, and 31.98% of the collections were rated higher by the state than the teacher.

Teacher Survey Results

The teachers who rated the audited writing collections were asked to complete a survey of the TOP training and rating process. A total of 458 surveys were submitted with the writing collections. The survey results indicated that approximately 93% of the raters felt that the TOP rubrics were sufficiently clear for use in assigning writing proficiency ratings. The results also indicated that 93% of the raters felt they had enough information about the writing abilities of the students to make judgments about their proficiency levels. The survey results indicated, however, fairly wide variation in the length of the rater training sessions and use of the state-provided training materials. While some raters received an in-depth training on the TOP rubrics and were given time for practice and discussion, others received little training. Approximately 25% of the raters reported that they did not receive training on two of the major state-provided training components intended to ground raters in the holistic rating rubric.

Conclusions/Limitations/Future Steps

It was originally hoped that this study would generate baseline rater agreement data that would help the state examine the accuracy of the TOP ratings. It was discovered, however, that this objective of the study could not be fully met given the TOP administration instructions to raters. TOP raters were instructed to rate students based on classroom observations and assemble writing collections that would support their ratings in the event of a state audit. In the state audit, however, only the students' writing collections were able to be evaluated. Therefore, when there was disagreement between the state and teacher ratings, it was not possible to know whether the teacher rated the student inaccurately (a holistic rating training issue) or whether the teacher's rating was accurate but the teacher failed to assemble a representative writing collection (an administration procedure issue).

While this first audit did not yield conclusive evidence of TOP rating accuracy, it was instrumental in helping the state determine ways to improve the TOP training and administration processes. Information from the study was shared with the state's technical advisory committee and English language learner focus group. The various recommendations both from the technical advisory committee and from the English language learner focus group are summarized as follows.

- Require a more uniform and in-depth training to help ensure that raters are adequately trained to apply the rating rubrics in a consistent manner across the state. Consider adding a qualifying component to the training system to help meet this objective.
- Modify the TOP administration directions such that raters base their writing ratings solely on the contents of student writing collections. This will allow districts and the state to better monitor rating reliability and validity.
- Provide more specific training and be more prescriptive concerning the number and types of writing samples to include in the writing collections. This will help ensure

that raters assemble collections that accurately portray the students' writing proficiency.

- Require that the components of the writing collections be verified for appropriateness by an independent reviewer on the campus before the rating process is complete.
- Conduct another writing study after the spring 2006 administration and include a substantially larger number of writing collections.

TEA plans to implement these recommendations for the spring 2006 TOP administration.

Table 1. 2005 TOP Sample vs. TOP Population

	Full TOP Dataset for Grades 2-12	TOP Study Sample (25 per Proficiency Level per Band)
% Male	53%	50%
% Hispanic	94%	91%
% in Bilingual Grades 2-5 Only	66%	58%
% in ESL Grades 2-5 Only	27%	33%
Number of Districts	1083	146
% Grade 2	20%	14%
% Grade 3	18%	11%
% Grade 4	12%	13%
% Grade 5	10%	12%
% Grade 6	9%	11%
% Grade 7	7%	8%
% Grade 8	6%	6%
% Grade 9	8%	13%
% Grade 10	4%	6%
% Grade 11	3%	3%
% Grade 12	2%	3%

Note: Grade bands used to select numbers of writing collections for this study are 2-3, 4-5, 6-8, 9-12.

Table 2. Distribution of Ratings by Teachers and by State

Rating	Teacher Rating		State Rating	
	N	%	N	%
Beginning	72	21%	36	10%
Intermediate	89	26%	108	31%
Advanced	91	26%	120	35%
Advanced High	95	27%	83	24%
Total	347	100%	347	100%

Table 3. Exact Agreement Rates by Grade Band

Grade Band	% Agreement
2-3	42%
4-5	45%
6-8	40%
9-12	42%
Overall	42%

Table 4. Exact Agreement Rate Between Teacher and State Ratings by Teacher Rating

Teacher Rating	% Agreement with State Rating
Beginning	38%
Intermediate	43%
Advanced	43%
Advanced High	44%
Overall	42%

Table 5. Relationship Between State Ratings and Teacher Ratings

Frequency Percent	State Rating B	State Rating I	State Rating A	State Rating H	Total Teacher Ratings
Teacher Rating B	27 7.78	29 8.36	10 2.88	6 1.73	72 20.75
Teacher Rating I	6 1.73	38 10.95	31 8.93	14 4.03	89 25.65
Teacher Rating A	1 0.29	30 8.65	39 11.24	21 6.05	91 26.22
Teacher Rating H	2 0.58	11 3.17	40 11.53	42 12.10	95 27.38
Total State Ratings	36 10.37	108 31.12	120 34.58	83 23.92	347 100.00

Exact Agreement (on the diagonal): 42.07%

Teacher writing rating higher than state rating (lower left corner): 25.95%

Teacher writing rating lower than state rating (upper right corner): 31.98%