Appendix 10

Texas Observation Protocols Audit Analysis

Grades 2–12 Writing Spring/Summer 2006

Background

The purpose of the Grades 2-12 writing component of the Texas Observation Protocols (TOP) audit is to evaluate the extent to which students' English language writing proficiency is rated congruent with state-defined procedures. An initial audit of TOP writing collections was conducted after the spring 2005 TOP administration. Based on information from the 2005 audit, requests from school districts, and input from the Texas Technical Advisory Committee (TTAC) and English Language Learner (ELL) Assessment Focus Group, the Texas Education Agency (TEA) and Pearson Educational Measurement (PEM) conducted another audit with several modifications following the spring 2006 TOP administration. The results of the 2005 study are summarized below.

In the 2005 TOP writing administration, teachers trained to be TOP raters were instructed to rate the English language learners (ELLs) in their classes based on observations and experience with the students and then to collect evidence to support their assigned ratings in the event of a state audit. During the audit, however, because of the observational nature of the assessment, it was difficult to isolate the source of disagreement between teacher ratings and the state's audit ratings. A particular student could have been rated incorrectly by the teacher (a training issue), or the rating may have been accurate but the rater may have failed to provide a writing collection with adequate evidence of the student's proficiency level (a documentation issue). In other words, part of the disagreement between teacher and state ratings could have been that the collections submitted were not truly representative of the students' proficiency.

The results of the 2005 study and feedback from the TTAC and an ELL Assessment Focus Group led TEA to make several changes in the TOP writing assessment procedures.

- Starting in the 2005-06 school year, teachers were instructed to base their ratings strictly on the collection of writing samples they gathered. This ensured that the teacher rater and the state rater were scoring the same thing.
- Writing collections in 2005 were required to contain at least three pieces of student writing. In 2006, writing collections were required to contain **at least five** pieces of student writing. Required procedures for verifying the correct assembly of the writing collection were also implemented for the 2006 administration.
- In addition, the state added a qualifying component to the assessment in 2006 to help ensure that teachers had received adequate training to appropriately assign ratings. In the 2005-06 school year, after practicing rating writing collections as part of their training, the raters also

rated writing collections online as the qualification component of their training. They had two attempts to qualify. If unsuccessful on the first set of collections, they rated a second set. Feedback on their ratings was provided after each set. To be successful, they had to rate correctly 7 out of 10 collections or 10 out of 15 collections depending on the number of grade levels they were trained to rate. About 54,000 teachers completed the qualification process and about 50,000 (92%) were successful.

Feedback from the ELL Assessment Focus Group led TEA to make additional changes in how the 2006 audit would be conducted. The focus group indicated the desire for districts to receive audit feedback to allow them to address any needed changes in TOP training sessions. The group also suggested that the audit be expanded to include more districts and campuses. As a result, the 2006 state audit underwent these changes:

- The audit sample of writing collections was much larger than in 2005 (total N = 14,000 in 2006 vs. N = 480 in 2005). Not only was the sample designed to be large enough to make sound statistical inferences about agreement rates between teachers and the state across both proficiency levels and grade cluster, it was designed to be of sufficient size to make inferences at the region and large-district level.
- Because of the increased sample size, districts and schools had a much higher probability of being selected for the audit.

Sampling Strategy

The sampling design used in 2006 differed from the random sampling method used in 2005. In 2006, sampling was conducted at a training "entity" level. This sampling plan was reviewed with the TTAC in February 2006.

In order to closely monitor rating efficacy, a plan was developed that involved sampling and providing feedback to the 20 regional educational service centers (ESCs) and the largest 35 ELL districts every year. Providing feedback at this level supports the validity and reliability of the assessment program while also satisfying region and district requests to receive feedback on the accuracy of their ratings. Through a training-of-trainers model, TEA and PEM delivered TOP training to these 55 "entities" who then delivered the training to teachers and other local campus and district personnel. By combining the 35 ELL districts and 20 service centers, 55 sampling units were created. Because in the 2005-06 school year the 35 largest ELL districts received TOP training separately from the ESC with which they are associated, a decision was made to consider these 35

districts as separate and mutually exclusive training entities. For example, although Houston is physically located within in the Region 4 ESC, it was sampled as a separate unit from Region 4, and the Region 4 unit does not include Houston. These considerations led to the implementation of the sampling plan as follows.

The plan is to audit all 55 entities (hence, all districts in the state) over a 2-3 year period, with 25 entities audited in year 1 (spring 2006). To select the 25 entities, the 55 entities were sorted by ELL enrollment, and 5 "strata" of 11 entities were created (see Table 1). Of the 25 training entities selected for the 2006 audit—7 entities from the first stratum, 5 from the second and third strata, and 4 from the fourth and fifth strata were randomly selected. The larger number of entities sampled from strata 1, 2, and 3 ensured that the sample was overall representative of the state's ELL population. The requirement to sample some entities from every stratum ensured that entities of varying size were represented. Approximately 560 students were then sampled from each of the 25 selected entities as follows.

As shown in Table 2, the 560 writing collections were comprised of 35 students from each of the four proficiency levels (Beginning, Intermediate, Advanced, and Advanced High) assigned by the teacher rater and each of the following grade clusters: 2, 3-5, 6-8, and 9-12. The sampling resulted in 560 * 25 = 14,000 student writing collections.

This sampling plan allowed for the large districts to drive the characteristics of the sample, while at the same time sampling the smaller units enough to provide unit-specific feedback. This plan will also allow for all 11 units in strata 4 and 5 (i.e., the smallest entities) to receive feedback within three years.

Table 1. Strata Created to Draw Sample

Unit	Stratum	Sampling Entity	Number of ELL Students	Region	% of State ELL pop	% of State ELL pop per Stratum
1	1	HOUSTON ISD	51301	4	8.34	
2	1	DALLAS ISD	43052	10	7.00	
3	1	REGION 4 (minus large districts)	31547	4	5.13	
4	1	REGION 1 (minus large districts)	31331	1	5.09	
5	1	BROWNSVILLE ISD	20841	1	3.39	
6	1	FORT WORTH ISD	19148	11	3.11	**46.04**
7	1	REGION 10 (minus large districts)	18499	10	3.01	
8	1	EL PASO ISD	17878	19	2.90	
9	1	REGION 11 (minus large districts)	17644	11	2.87	
10	1	REGION 20 (minus large districts)	16115	20	2.62	
11	1	AUSTIN ISD	15955	13	2.59	
12	2	REGION 13 (minus large districts)	15685	13	2.55	
13	2	LAREDO ISD	14243	1	2.31	
14	2	UNITED ISD	13597	1	2.21	
15	2	ALIEF ISD	13395	4	2.18	
16	2	ALDINE ISD	13299	4	2.16	
17	2	REGION 7	12097	7	1.97	**22.25**
18	2	GARLAND ISD	11285	10	1.83	
19	2	PASADENA ISD	11216	4	1.82	
20	2	REGION 6	10945	6	1.78	
21	2	YSLETA ISD	10940	19	1.78	
22	2	PHARR-SAN JUAN-ALAMO ISD	10252	1	1.67	
23	3	CYPRESS-FAIRBANKS ISD	9889	4	1.61	
24	3	LA JOYA ISD	9825	1	1.60	
25	3	REGION 19 (minus large districts)	9569	19	1.55	
26	3	IRVING ISD	9499	10	1.54	
27	3	SOCORRO ISD	9294	19	1.51	**15.85**
28	3	ARLINGTON ISD	9057	11	1.47	
29	3	SPRING BRANCH ISD	8585	4	1.39	
30	3	SAN ANTONIO ISD	8308	20	1.35	
31	3	REGION 12	8082	12	1.31	
32	3	MCALLEN ISD	7840	1	1.27	
33	3	REGION 18	7611	18	1.24	
34	4	EDINBURG CISD	7235	1	1.18	
35	4	REGION 16	6448	16	1.05	
36	4	FORT BEND ISD	6062	4	0.99	
37	4	REGION 2	5969	2	0.97	
38	4	RICHARDSON ISD	5712	10	0.93	**10.16**
39	4	DONNA ISD	5603	1	0.91	
40	4	CARROLLTON-FARMERS BRANCH	5513	10	0.90	
41	4	PLANO ISD	5490	10	0.89	
42	4	GALENA PARK ISD	5073	4	0.82	
43	4	EAGLE PASS ISD	4804	20	0.78	
44	4	RIO GRANDE CITY	4607	1	0.75	
45	5	KATY ISD	4278	4	0.70	
46	5	NORTHSIDE ISD	4271	20	0.69	
47	5	GRAND PRAIRIE ISD	4153	10	0.67	
48	5	LEWISVILLE ISD	4063	11	0.66	
49	5	REGION 17	3623	17	0.59	
50	5	REGION 15	3583	15	0.58	**5.70**
51	5	REGION 8	3351	8	0.54	
52	5	REGION 5	2720	5	0.44	
53	5	REGION 3	2494	3	0.41	
54	5	REGION 14	1292	14	0.21	
55	5	REGION 9	1255	9	0.20	

Table 2. Number of Students Sampled for a Given Entity

Grade Cluster	Beginning	Intermediate	Advanced	Advanced High	Total
2	35	35	35	35	140
3-5	35	35	35	35	140
6-8	35	35	35	35	140
9-12	35	35	35	35	140
Total	140	140	140	140	**560**

Sample Characteristics

The above sampling procedure resulted in selection of the following entities for inclusion in the 2006 audit.

Table 3. Entities Selected for 2006

Entity
Houston ISD
Dallas ISD
Region 4 minus large districts
Region 1 minus large districts
Brownsville ISD
Fort Worth ISD
Region 10 minus large districts
Region 13 minus Austin ISD
Laredo ISD
Alief ISD
Pasadena ISD
Pharr-San Juan-Alamo ISD
Region 19 minus large districts
Spring Branch ISD
San Antonio ISD
McAllen ISD
Region 18
Edinburg CISD
Region 16
Region 2
Rio Grande City ISD
Northside ISD
Region 8
Region 5
Region 3

The characteristics of the 2006 sample closely matched that of the TELPAS testing population. Table 4 below indicates how the sample compares to the population.

	Full Grades 2-12 Spring 2006 TELPAS Population	**TOP Writing Sample**
Number of K-12	639,829	--
Number of 2-12 Only	449,404	13,990*
% Male	53%	52%
% Hispanic	93%	94%
% Grade 2	20%	25%
% Grade 3	19%	11%
% Grade 4	12%	8%
% Grade 5	11%	6%
% Grade 6	7%	8%
% Grade 7	7%	9%
% Grade 8	6%	8%
% Grade 9	8%	11%
% Grade 10	4%	6%
% Grade 11	3%	4%
% Grade 12	2%	3%
% ESL (2-5 Only)	**27%**	**24%**
% Bilingual (2-5 Only)	**66%**	**67%**

Table 4. Comparison of 2006 Sample and TELPAS Testing Population

**Please note that the desired total sample size was 14,000. However, because some of the smaller entities did not have enough students in each proficiency level to sample, the total sample size was 13,968.*

Scoring Process

The Performance Scoring Center rescored the audited writing collections using a two-phase process.

Phase I: Initial Scoring

During the first phase, the state raters were trained to rate collections using the same rubrics and training sets used to train the district raters within each entity.

After training, the state raters scored the writing collections following the same process used for other state assessments. Each writing collection was scored independently by two state raters without knowledge of the original district rating. If the two state ratings matched, the Phase I rating was recorded and processed. In cases where the first two state scores did not match, the collection

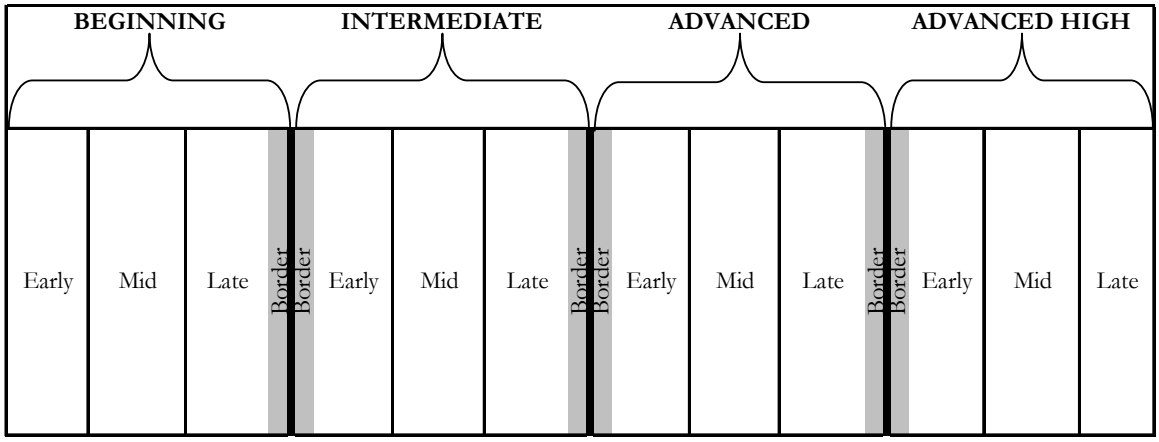
was read by a third scorer. Then, if two of three raters agreed, this rating was recorded and processed. If disagreement was still present, the lead scoring director resolved the rating. Throughout this process, periodic checks were made to ensure against rater drift.

Phase II: Analytic Scoring of Collections

In Phase II, the state trained teams of raters to analytically rescore collections for which the entity and state ratings differed by one proficiency level. The purpose of Phase II scoring was twofold. Fist, it enabled the state to verify the Phase I scoring since this was the first large-scale state audit, and, although the state's raters were experienced scorers on other projects, they did not have previous experience with TOP. The second purpose was to learn more about the reasons for the adjacent ratings in order to identify writing collections that would be useful for TOP training materials in the future.

The training conducted in this phase prepared the analytic scorers to identify collections as Early, Mid, or Late within a proficiency level range or as Border (near the line between two proficiency levels). The training increased the expertise of the state's scorers and led them to reconsider a number of Phase I scores. In Phase II scoring, the collections were assigned not only a proficiency level score but also a designation of Border, Early, Mid, or Late as shown in Figure 1.

Figure 1. Depiction of Phase II Classifications



The analytic raters provided one additional rating, working collaboratively with other analytic raters and lead state trainers to determine the final rating. All collections receiving Border designations were subsequently rated by the lead state trainer from PEM or by a PEM scoring director.

The collections rated Border during this phase were assigned the district rater's score. Ratings for these collections were deemed too close to dispute for purposes of the audit. In addition, the state ratings from Phase II (which reflected additional, more targeted training) overrode the state ratings from Phase I ratings if they differed. The Early, Mid, and Late sublevel designations will enable the state to identify collections to fulfill future training needs.

Results

Of the 13,968 student writing collections requested, the school districts submitted 13,902 (99.5%) collections. Of the 13,902 received collections, 293 (2.1%) were classified as nonscorable by the scoring center. Collections were deemed nonscorable if they could not be used to judge the overall writing proficiency of a student because of an insufficient number, variety, or type of writing samples. Also, 252 (1.8%) collections were received for students served by special education who were noted to be at IEP instructional level K or 1. These collections were not rescored in this audit because the state raters were trained to only score grades 2-12 writing collections. A total of 13,357 writing collections were re-rated by the state.

Table 5 shows the distribution of collection ratings by teachers and the state. Although the sample was designed to obtain equal numbers of collections across grade clusters and teacher-assigned proficiency level, the score distributions resulting from the state's rescoring were slightly different.

Table 5. Distribution of Ratings by Entities and by State

	Ratings by Entities		Ratings by State	
	N	%	N	%
Beginning	3059	23	2218	17
Intermediate	3420	26	3861	29
Advanced	3445	26	4105	31
Advanced High	3433	26	3173	24
Total	**13357**	**101***	**13357**	**101***

*Percents do not add exactly to 100% due to rounding.

Tables 6, 7, and 8 present the adjacent, modified adjacent, and perfect agreement rates (respectively) between the state and teacher ratings. As seen in table 7, the state ratings agreed perfectly or adjacently (plus or minus 1 level) with the teacher ratings 98% of the time. The modified adjacent agreement rates were computed by considering state ratings in agreement with teacher ratings if they either 1) perfectly agreed or 2) were rated by the state as Early or Late within a proficiency level

adjacent to the rating provided by the teacher. As seen in table 8, the modified agreement rate was 90%. Finally, table 8 shows that the state ratings agreed perfectly with the teacher ratings 77% of the time. There was little variation in the agreement rates across grade clusters, suggesting that enrolled grade level had little or no impact on rating accuracy.

Table 6. Adjacent Agreement Rates Between Entities and State by Grade Cluster

Grade Cluster	% Agreement
2	98
3-5	98
6-8	98
9-12	97
Overall	**98**

Table 7. Modified Adjacent Agreement Rates Between Entities and State by Grade Cluster

Grade Cluster	% Agreement
2	90
3-5	89
6-8	90
9-12	89
Overall	**90**

Table 8. Perfect Agreement Rates Between Entities and State by Grade Cluster

Grade Cluster	% Agreement
2	78
3-5	78
6-8	78
9-12	76
Overall	**77**

The minimum standard for perfect agreement between two raters of a performance scored item or assessment depends upon 3 factors:

1) the subject area of the assessment,
2) the type of rubric (holistic vs. analytic), and
3) the number of points on the rubric.

For a writing assessment using a 4-point holistic rubric, this standard is 70% perfect agreement. (source: PEM PSC ISO Standards)

Table 9 presents the perfect agreement rates between the state and teacher ratings grouped by the teacher rating. The perfect agreement rate was lowest (69%) for writing collections rated as beginning by the teacher, and highest (81%) for those rated as advanced. The perfect agreement rates for intermediate (78%) and advanced high (79%) were right around the overall state average.

Table 9. Perfect Agreement Rate Between Entities and State by Proficiency Level

Entity Rating	% Agreement with State Rating
Beginning	69
Intermediate	78
Advanced	81
Advanced High	79
Overall	**77**

Table 10 presents a cross-tabulation of teacher rating by state rating. In addition to the 77% of the collections that received the same rating by the teacher and the state, this analysis indicates that 15% of the collections were rated higher by the state, and 8% of the collections were rated higher by the teacher rater.

Finally, two summary indices were calculated to further evaluate the degree of association between the teacher and state ratings. The teacher and state ratings were found to be very highly correlated ($r = 0.87$, $p < 0.001$). In addition, a weighted kappa statistic, K_w, was calculated. The ***K*** value can be interpreted as follows (Altman, 1991):

Value of ***K***	Strength of agreement
< 0.20	Poor
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Good
0.81 - 1.00	Very good

The obtained K_w from this study was 0.7882, indicating "good" relationship between teacher and state ratings.

Table 10. Relationship between State Ratings and Teacher Ratings

Frequency Percent	State Rating B	State Rating I	State Rating A	State Rating H	Total Teacher Rating
Teacher Rating B	2126 15.92	818 6.12	96 0.72	19 0.14	3059 22.90
Teacher Rating I	69 0.52	2681 20.07	592 4.43	78 0.58	3420 25.60
Teacher Rating A	18 0.13	261 1.95	2807 21.01	359 2.69	3445 25.79
Teacher Rating H	5 0.04	101 0.76	610 4.57	2717 20.34	3433 25.70
Total State Ratings	2218 16.61	3861 28.91	4105 30.73	3173 23.76	13357 100.01**

***Total percentage exceeds 100 due to rounding.*

Perfect Agreement (on the diagonal): 77.34%
Teacher writing rating higher than state rating (lower left corner): 7.72%
Teacher writing rating lower than state rating (upper right corner): 14.69%

Survey Results

Two surveys were given as part of the audit study. One was given to district and campus testing coordinators, and the other was given to teachers who rated the audited writing collections.

Results of Testing Coordinator Survey – This survey was used to gather information about procedures the audited districts followed to support the validity and reliability of the TOP assessment process. District testing coordinators, as well as campus testing coordinators in the audited schools, were required to complete the survey. Across the 25 audited entities, 95% of the district coordinator responses indicated that *district-wide* procedures were implemented to support the validity and reliability of the assessment process. Table 11 displays the types of procedures districts reported using. Districts were directed to mark all procedures that applied.

Table 11. Testing Coordinator Survey Results

Procedures Used to Support Reliability and Validity of TOP Ratings	Percent	Number
Backup qualified raters provided second ratings of collections rated by nonqualified raters	45%	166
Any nonqualified raters collaborated with qualified raters	54%	199
Any nonqualified raters collaborated with a highly qualified, lead rater	26%	96
Qualified raters collaborated with other qualified raters as needed	85%	312
Qualified raters collaborated with a highly qualified, lead rater as needed	42%	152
Writing collections were randomly selected and rated a second time	18%	67
All writing collections were rated a second time	20%	74
Resolution ratings were provided when ratings given by two or more raters differed	23%	83

Of the 5% responding that they did not implement *district-wide* procedures, 37% responded that some of their campuses implemented these types of procedures as a campus option.

Results of TOP Rater Survey – The rater survey was designed to gather information about the extent to which raters felt prepared to assess their students' proficiency levels using the TOP scoring rubrics. On an audited campus, only raters whose students' writing collections were selected for the audit were directed to complete the survey. More than 6,000 rater surveys were submitted. Across the 25 audited entities, 93% of the responses indicated that the training was sufficient to prepare raters for the online qualification portion of their training. The responses indicated that 92% of the raters completed the online qualification component successfully. Table 12 below shows the percent of responses by language domain of raters that felt they had sufficient information to judge the English language proficiency levels of their students.

Table 12. TOP Rater Survey Results

Language Domain	Percent
Listening, grades 2-12	96%
Speaking, grades 2-12	96%
Writing, grades 2-12	97%
Reading (grade 2 only*)	93%

***RPTE assesses the reading domain in grades 3-12.**

These survey results indicate that a vast majority of districts and campuses implemented a procedure that supported the rating process through rater collaboration, consultation with a lead rater, or rating verification. In addition, the vast majority of raters reported that they had received adequate training,

were successful on the online qualification component for grades 2–12 writing, and had adequate information to rate their students' proficiency levels in each of the language domains assessed.

Conclusions/Next Steps

In 2006 the state implemented several steps to improve the audit process so that rater agreement data could be appropriately assessed and attributed. It appears that the changes were beneficial. The results of the 2006 audit were positive. The overall perfect agreement rate of 77% exceeded the state's 2005-2006 expectation based on the PEM PSC ISO standard, and the adjacent agreement rate was 98%. In addition, the presence of a high correlation and a high weighted kappa value underscored the strong agreement between the state ratings and the teacher ratings. The observed agreement rates did not vary much across the audited entities, grade cluster, or proficiency level.

The audit revealed no significant problems with rating accuracy. TEA and PEM will continue the annual audit process to provide ongoing monitoring of rating effectiveness and to give training entities feedback and enhanced training materials that can refine the ability of raters to conduct this assessment. The expectation is that rating accuracy will continue to improve over the next several years as raters become increasingly familiar with the rating process.