TELPAS Writing Audit Report

Grades 2–12
Spring/Summer 2019

Introduction

Overview

The purpose of the Texas English Language Proficiency Assessment System (TELPAS) writing audit is to provide ongoing evidence of the validity and reliability of the holistically rated writing component of TELPAS. Under this assessment system, teachers are trained to use the TELPAS Proficiency Level Descriptors (PLDs) to rate the English language proficiency of students in specified language domains, including writing. The 2019 primary audit activities included:

1. evaluating the extent to which the writing English language proficiency ratings assigned by teacher raters match those assigned by state audit raters,
2. gathering feedback from the teacher raters on the quality of their training for writing, and
3. examining how well educators followed state-defined administration procedures to rate writing performance.

Seven writing audits (including the 2019 audit) have occurred to date. The first audit, conducted in 2005, was relatively small, allowing the state to examine and improve audit procedures. The second audit, conducted in 2006, was a larger study in which information from a large, representative number of districts and students was collected. That audit provided results for regional education service centers (ESCs) and large districts referred to as *training entities* because of their roles in directly providing TELPAS training to teacher raters. The 2006 audit results provided reliability and validity evidence supporting the accuracy of teacher ratings on the writing domain. The 2007, 2008, 2013, 2016, and 2019 TELPAS audits were smaller audits and served to provide ongoing evidence of validity and reliability at the level of the state rather than at the training entity level. The comparison of rater agreement across the five audits indicated that rater accuracy improved from 2005 to 2006, and was stable across 2006, 2007, 2008, and 2013. In 2016, there was a slight increase in rater accuracy from 2013. In 2019, there was a relatively large increase in agreement rates from previous audits. This report includes the sampling plan, the audit results, and questionnaire results for the 2019 writing audit.

Sampling Strategy

For the 2019 writing audit, students were randomly sampled in grades 2–12 from across the state, including all 20 regions. To minimize testing burden on Texas schools, the following types of schools were excluded from the sample:

- Campuses that had been rated as Improvement Required (IR) according to the state accountability system for three or more years
- Alternative education campuses (including juvenile justice alternative education programs and disciplinary alternative education programs)
- Campuses selected for the 2016 TELPAS Writing Audit

The target sample size for the 2019 writing audit was 2,000 students, spread equally across grade bands and proficiency levels to allow for analysis by grade band and by proficiency level. This target sample size allowed for a representative sample of student writing collections while minimizing the time required for completing the audit task. Given that some schools might not be able to provide writing samples for all students selected for the target sample, an extra 10% was added to the target sample to ensure that at least 2,000 students were included. As shown in Table 1, 138 students were randomly sampled from each of the four proficiency levels (beginning, intermediate, advanced, and advanced high) assigned by the teacher rater and from each of the following grade bands: 2, 3–5, 6–8, and 9–12. The target sample included 2,208 students.

Table 1. Total Number of Students in Target Sample

Grade Bands	Beginning	Intermediate	Advanced	Advanced High	Total
2	138	138	138	138	552
3–5	138	138	138	138	552
6–8	138	138	138	138	552
9–12	138	138	138	138	552
Total	**552**	**552**	**552**	**552**	**2,208**

Sample Characteristics

The 2019 target sample and tested population is compared in Table 2. The 2019 target sample was representative of the population of students who took TELPAS in grades 2–12 in 2019 in terms of gender and ethnic representation, as well as representative of the percent of students in grades 2–12. The sample was less representative of the population in terms of the percent of students in English as a Second Language (ESL) and bilingual programs, due in part to the slightly lower percentage of elementary students in the sample compared to the population, and the slightly higher percentage of middle school and high school students in the sample compared to the population. In Texas, bilingual programs are more common in elementary schools, whereas ESL programs are more common in middle and high schools.

Table 2. Comparison of the 2019 Population and Audit Target Sample

Student Characteristics	2019 Population	2019 Audit Sample
Number of 2–12 Students	770,430	2,208
% Male	53.4%	55.1%
% Hispanic	90.0%	89.2%
% Grade 2	13.4%	25.0%
% Grade 3	13.5%	9.3%
% Grade 4	13.1%	8.3%
% Grade 5	11.9%	7.4%
% Grade 6	10.4%	10.0%
% Grade 7	9.0%	8.7%
% Grade 8	8.0%	6.3%
% Grade 9	7.8%	10.8%
% Grade 10	5.5%	6.1%
% Grade 11	4.2%	4.8%
% Grade 12	3.3%	3.3%
% ESL	60.7%	61.0%
% Bilingual	34.9%	34.4%

Rescoring of Writing Collections

The TELPAS writing assessment for grades 2–12 required teacher raters to assemble a collection of each English language learner's writing from classroom instruction in a variety of core content areas. After the scores were collected from the teacher raters, the Pearson Performance Scoring Center (PSC) rescored the writing collections using a two-phase process described below.

Phase I: Initial Scoring

As part of the first phase, audit raters were trained to rate collections using the same rubrics and training sets used to train teacher raters. In order to qualify, the audit raters were required to rate correctly at least 80% of the writing collections from a qualifying set, a set of 15 writing collections which had been previously rated. There were two qualifying sets. If an audit rater did not qualify on the first qualifying set administered, he or she needed to qualify on the second qualifying set to participate in the audit study.

After audit raters were trained and qualified, each writing collection was scored independently by two audit raters without knowledge of the teacher rating. If the two audit ratings matched, the Phase I rating was recorded. In cases where the audit raters did not match, the collection was scored by a third audit rater. Then, if two of three audit raters agreed, this rating was recorded and processed. If disagreement was still present, a highly qualified senior rater resolved the rating. Throughout this process, periodic checks were made to help prevent rater drift among the audit raters, including rescoring of approximately 5% of Phase 1 ratings by trained scoring supervisors to identify any issues with rater drift. If rater drift was identified through these periodic checks, feedback was provided to the raters to correct the issue.

Phase II: Analytic Scoring of Collections

Multiple purposes motivated Phase II of the scoring process. First, a second phase enabled the state to verify Phase I scores where the ratings from the teachers and audit raters were different. Second, the results from Phase II enabled the state to learn more about the reasons for adjacent ratings. Finally, the state was able to identify writing collections that could be useful for future training materials.

In Phase II, the trained team of scoring supervisors, who supervised the Phase I audit raters, rescored all writing collections for which the teacher and audit ratings differed by at least one proficiency level. Collections were read aloud to the team of scoring supervisors to avoid any bias that could be created due to phonetic spelling of words. The team did not know the Phase I audit rating nor the teacher rating, only that there was a difference between the two ratings. After hearing each student's collection read aloud, the team reached consensus about the proficiency level of the student. The ratings from Phase II (which reflected additional, more targeted training) overrode the ratings from Phase I if they differed.

Final ratings from Phase I, where collections did not need a Phase II review, and final ratings from Phase II together make up the full set of audit rating results. These ratings will be referred to throughout the remainder of the report as *final audit ratings*.

2019 Audit Results

Of the 2,208 student writing collections requested, sampled schools submitted 2,203 (99.8%) collections. Of the 2,203 collections received, all of them were classified as scorable 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. Consequently, a total of 2,203 writing collections were re-rated by the state.

Table 3. Distribution of Ratings by Teachers and Audit Raters

Proficiency Level	Ratings by Teachers		Ratings by Audit Raters	
	N	%	N	%
Beginning	548	25	514	23
Intermediate	551	25	558	25
Advanced	554	25	613	28
Advanced High	550	25	518	24
Total	**2,203**	**100**	**2,203**	**100**

The distribution of ratings by teachers and the audit raters is provided in Table 3. The sample was designed to obtain equal numbers of collections across grade bands using teacher-assigned proficiency levels. Although this objective was achieved, disagreements between teacher and audit ratings resulted in more writing collections classified into the advanced proficiency level and less into the beginning and advanced high proficiency levels by audit raters.

Tables 4 and 5 include the perfect and adjacent agreement rates between the teacher and audit ratings, respectively. The teacher rating was considered to have perfect agreement with the audit rating if the teacher rating was the same as the final audit rating. The final audit ratings agreed perfectly with the teacher ratings 90% of the time overall (refer to Table 4).

Table 4. Perfect Agreement Rates between Teacher and Audit Raters by Grade Band

Grade Band	% Agreement
2	91
3–5	93
6–8	89
9–12	86
Overall	**90**

The teacher and audit ratings were considered adjacent if the teacher rating was the same as or within one proficiency level of the final audit rating. The adjacent agreement rate was 99% overall (refer to Table 5).

Table 5. Adjacent Agreement Rates between Teacher and Audit Raters by Grade Band

Grade Band	% Agreement
2	99
3–5	99
6–8	99
9–12	99
Overall	**99**

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

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

For a writing assessment using a four-point holistic rubric, the standard is 70% perfect agreement (*Pearson Performance Scoring Center International Organization for Standardization (ISO) Standards*). The perfect agreement rates in Table 4 indicate that the TELPAS writing audit results exceeded the applicable standard.

Perfect agreement rates between the teacher and audit ratings are shown by proficiency level in Table 6. Perfect agreement was lowest (87%) for writing collections rated as Intermediate by the teacher and highest (92%) for those rated as Advanced.

Table 6. Perfect Agreement Rates between Teacher and Audit Raters by Proficiency Level

Teacher Rating	% Agreement with Audit Rating
Beginning	91
Intermediate	87
Advanced	92
Advanced High	89
Overall	**90**

A cross-tabulation of teacher by audit ratings is provided in Table 7. The diagonal highlighted cells indicate perfect agreement. The sum of these highlighted cells is 90% or the overall perfect agreement rate previously provided in Tables 4 and 6. Table 7 also includes the frequencies of adjacent and non-adjacent ratings (non-highlighted cells). Cells above the highlighted cells indicate that the collections were rated higher by the audit raters than by the teachers. Cells below the highlighted cells indicate that the collections were rated lower by the audit raters than by the teachers. Of the collections that did not receive the same rating from the teachers and audit raters, 6% of the collections were rated higher by the audit rater and 4% of the collections were rated lower by the audit rater.

Table 7. Relationship between Teacher and Audit Ratings

Frequency Percent	Audit Rating **Beginning**	Audit Rating **Intermediate**	Audit Rating **Advanced**	Audit Rating **Advanced High**	Total Teacher Rating
Teacher Rating **Beginning**	496 22.5%	47 2.1%	5 0.2%	0 0.0%	548 24.9%
Teacher Rating **Intermediate**	14 0.6%	481 21.8%	49 2.2%	7 0.3%	551 25.0%
Teacher Rating **Advanced**	3 0.1%	21 1.0%	510 23.2%	20 0.9%	554 25.2%
Teacher Rating **Advanced High**	1 0.1%	9 0.4%	49 2.2%	491 22.3%	550 25.0%
Total Audit Ratings	514 23.3%	558 25.3%	613 27.8%	504 23.2%	2203 100%

Finally, two indices were calculated to evaluate further the degree of association between the teacher and audit ratings, inter-rater reliability and weighted Kappa (refer to Table 8). Inter-rater reliability is the correlation between the teacher and audit ratings. The teacher and audit ratings were found to be highly correlated ($r = 0.943$, $p < 0.001$). Quadratic weighted Kappa (K_w) also provides a measure of inter-rater agreement for categorical ratings. The K_w value can be interpreted using the criteria in Table 9 (Altman, 1991). The overall K_w from this study was 0.908 indicating a *very good* relationship between teacher and audit ratings.

Table 8. Inter-Rater Reliability and Weighted Kappa between Teacher and Audit Ratings

Grade Band	Inter-Rater Reliability	Weighted Kappa
2	0.947	0.914
3–5	0.960	0.908
6–8	0.945	0.906
9–12	0.925	0.872
Overall	**0.943**	**0.908**

Table 9. Interpreting Kappa (K_w) Statistics

Value of K_w	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

Questionnaire Results

As a part of the 2019 audit, questionnaires were given to testing coordinators from the selected districts and campuses as well as to the teachers who had rated the audited writing collections. Questionnaires were used to gather information about the training and qualification procedures in order to support the validity and reliability of the TELPAS process. A brief description of the questionnaire results is provided below. The complete set of questions and responses is provided in the Appendix.

The TELPAS online basic training courses are provided to teach raters the essentials of second language acquisition theory. They also teach raters how to use the PLDs from the English Language Proficiency Standards (ELPS) to accurately identify the English language proficiency levels of their ELs based on how well the students understand and use English during daily academic instruction and classroom interactions. The online writing courses for grades 2–12 contain practice rating activities that comprise student writing collections. Online courses for K–1 contain numerous practice rating activities that comprise student writing samples and video segments in which ELs demonstrate their writing, reading, speaking, and listening skills in authentic Texas classroom settings. The courses give teachers practice applying the scoring rubrics (i.e., the PLDs) and provide teachers with detailed feedback about their rating accuracy.

Each year, raters are required to complete online calibration activities to demonstrate their ability to apply the scoring rubrics consistently and accurately before they rate students for the operational assessment. In 2018–2019, the calibration activities were provided for all domains for K–1 and for writing for grades 2–12. In addition, calibration activities were provided for raters of ELs approved for a special holistic administration for listening and speaking. There are two sets of calibration activities, and all applicable language domains are represented. In order to demonstrate sufficient calibration, raters are required to rate at least 7 out of 10 students correctly

within a set. Raters finish the calibration activities when they demonstrate sufficient accuracy. If sufficient accuracy is not obtained on the first set, the rater attempts a second and final calibration set. Individuals not successful on the final set are either not used as raters (a district decision) or are provided rater support in accordance with test administration procedures.
In general, testing coordinators reported that proper procedures are in place to train raters and evaluate proper implementation of the holistic ratings. Specifically, most district testing coordinators (DTCs) reported using many of the training materials provided by the state for teaching campus coordinators about administration procedures for the TELPAS holistic ratings. Additionally, 96% of DTCs reported implementing the required validity and reliability procedures including collaborative ratings and double rating of writing collections. Most districts (95%) selected the general procedures used to provide validity and reliability evidence for the TELPAS ratings, and many required campuses to keep documentation of the implemented procedures. The vast majority of DTCs (99%) reported that raters in their districts either passed the online training calibration by the end of the second calibration set or were provided with supplemental support. These results are based on a wide variety of districts, including districts that had between 1 and 10 raters trained (8.4%), and districts that had over 300 raters trained (17.2%).

The campus testing coordinator (CTC) responses were similar to the DTC responses. They reported using the state's training materials to provide rater training. Of the CTCs responding, 38% reported that all raters at their campuses had successfully calibrated by the end of the second calibration set. For those who were not successful calibrating, very few CTCs (1%) reported not providing supplemental rating support. More than half of the CTCs (62%) reported that individuals that were not successful in calibrating served as raters in 2019; of those that used uncalibrated raters, 18% had three or more. The CTCs reported a wide range of campus sizes; 16% of campuses had 1–10 raters, while 35% had more than 26.

Teacher raters reported feeling sufficiently prepared to provide holistic ratings. Teachers reported that 93% of all rater training sessions lasted between 30 minutes and 4 hours. Most new raters (85%) found the online basic training course adequate to prepare them for student ratings. Nearly all raters (92%) indicated that the calibration activities provided adequate preparation for their TELPAS ratings. On calibration set 1, 76% of teachers successfully calibrated; on calibration set 2, another 16% successfully calibrated. Of the raters who were unsuccessful at calibrating, 65% reported receiving rating support, as required by the state. Raters included teachers spanning grades K–12, including special education (3%), gifted and talented (5%), bilingual (30%), and ESL (25%) teachers from all content areas. Most teachers rated 25 or fewer students (75%), while only 7% of teachers reported rating 50 or more students.

In general, the questionnaire results provided additional evidence that the audit sample included a wide variety of districts, campuses, and teachers from around the state of Texas. Questionnaire responses also indicated that the vast majority of DTCs, CTCs, and teacher raters are implementing the TELPAS rater training and holistic rating procedures as described in the *2019 District and Campus Coordinator Resources* with fidelity.

Conclusions/Next Steps

The results of the 2019 audit provided evidence of rater accuracy at a higher level than what was reported in the previous audits. The overall perfect agreement rate of 90% was found to be satisfactory based on the *Pearson Performance Scoring Center ISO Standards*, and the adjacent agreement rate was 99%. In addition, the high correlation ($r = 0.943$) and high weighted Kappa value ($K_w = 0.908$) underscored the strong agreement between the teacher ratings and the audit ratings.

Furthermore, consistently high rating accuracy across five consecutive audits (perfect agreement rates were 76% in 2007, 79% in 2008, 77% in 2013, 81% in 2016, and 90% in 2019) provides evidence of growth in inter-rater agreement over time. Since TELPAS scores across years are used in reporting student progress in language acquisition, the state's finding of rater accuracy over time supports inferences about annual student progress from TELPAS scores.

Finally, the audit questionnaire results indicated that the vast majority of testing coordinators are implementing the training and ratings procedures as intended by the state. The DTC questionnaire results indicated that 96% of districts implemented procedures to support the validity and reliability of the TELPAS rating process. The CTC questionnaire results indicated that campuses were appropriately implementing the supplemental support activities for raters who were unsuccessful after their second calibration attempt and that they were implementing the validity and reliability procedures for the TELPAS rating process. Questionnaire results also indicated that the vast majority of raters reported that they agreed or strongly agreed to receiving adequate training if they were a new rater (85%), were successful on the online calibration component of rater training for grades 2–12 (92%), and had enough information to rate their students' proficiency levels in writing (96%).

The audit revealed no significant problems with implementation of training procedures or rater accuracy. The state plans to continue the writing audit process to provide ongoing monitoring of teacher rater effectiveness and to give district personnel feedback and training materials that best support the ability of raters to conduct this assessment.

References

Altman, D. G. (1991). *Practical Statistics for Medical Research.* London, UK: Chapman and Hall/CRC Press.

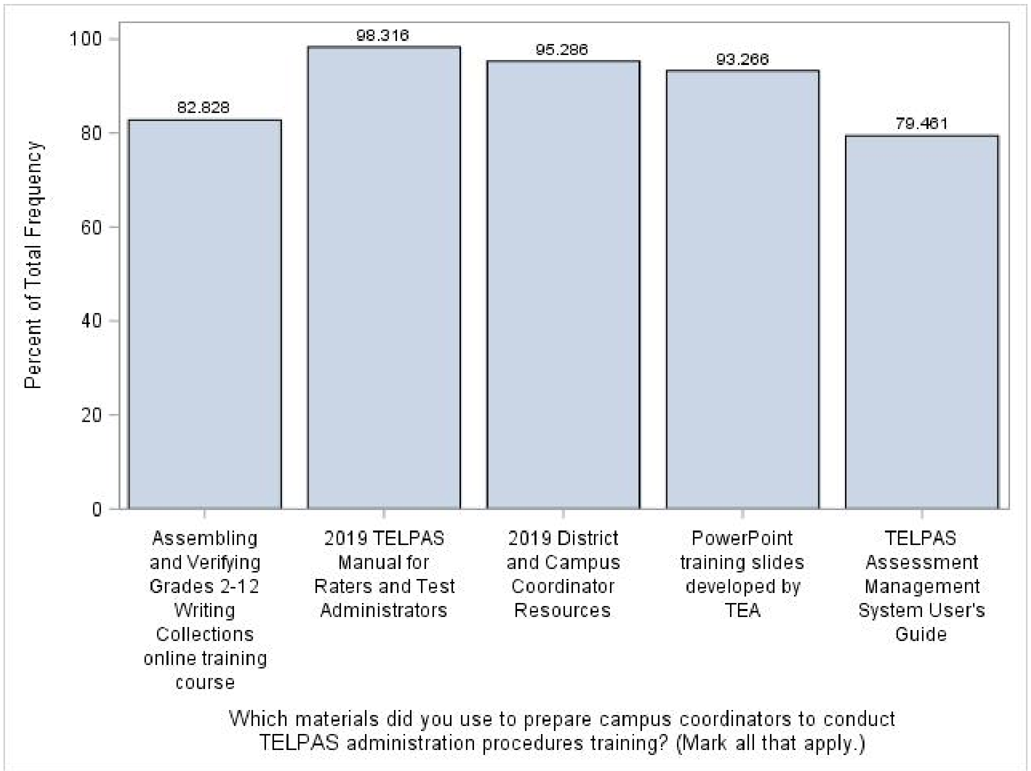
Appendix: 2019 Writing Audit Questionnaire Results

Questionnaire results are provided below for district testing coordinators, campus testing coordinators, and teacher raters. The questions and answers provided are for the multiple-choice items only.

District Testing Coordinators

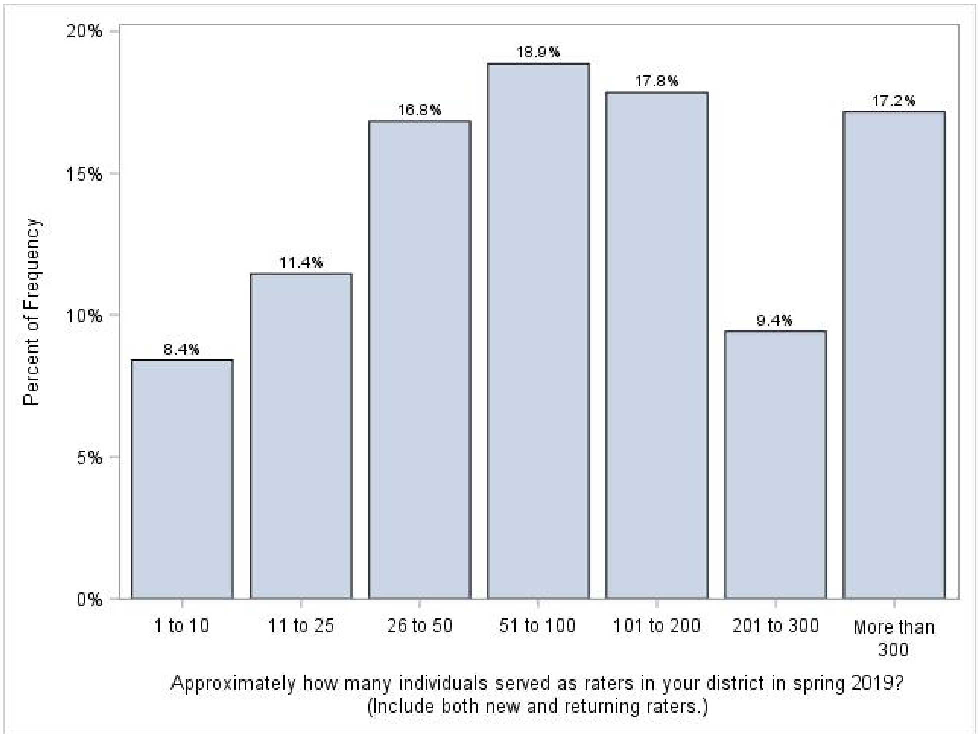
District testing coordinators (DTCs) at districts selected for participation in the 2019 TELPAS writing audit were asked to respond to a set of 12 questions. The first question asked them to select their district name. Questions 2–12 are provided below, along with their responses. In general, DTCs reported that proper procedures are in place to train raters and evaluate proper implementation of the holistic ratings. A total of 297 responses were received from DTCs.

2. What materials did you use to prepare campus coordinators to conduct TELPAS administration procedures training? (Mark all that apply)

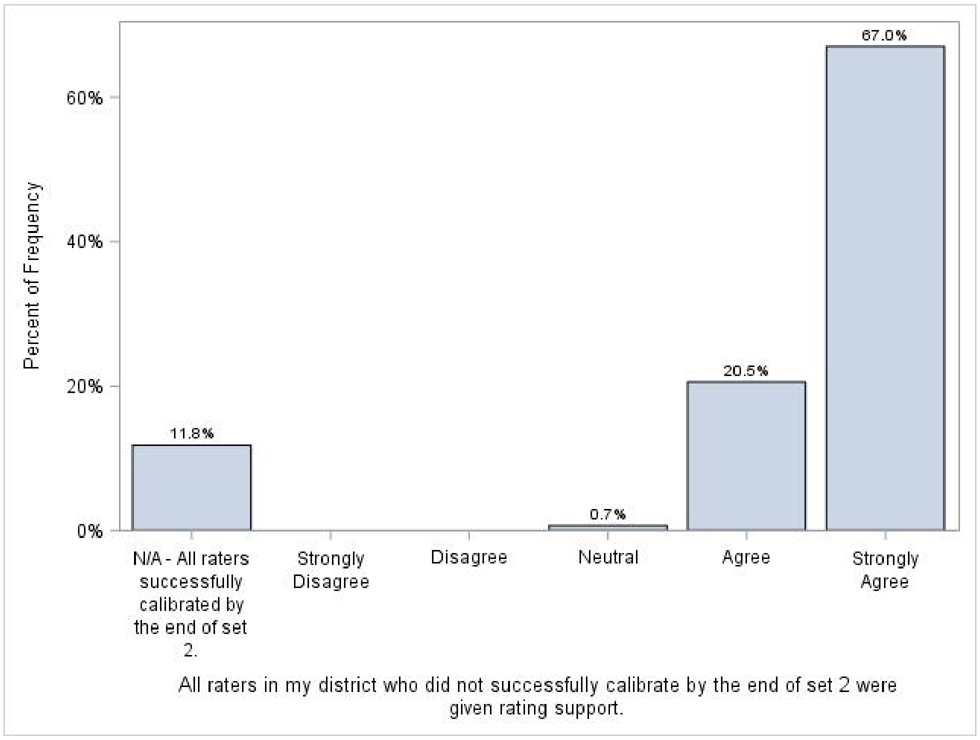


Note: Respondents were allowed to select more than one category. Therefore, totals do not necessarily equal 100%.

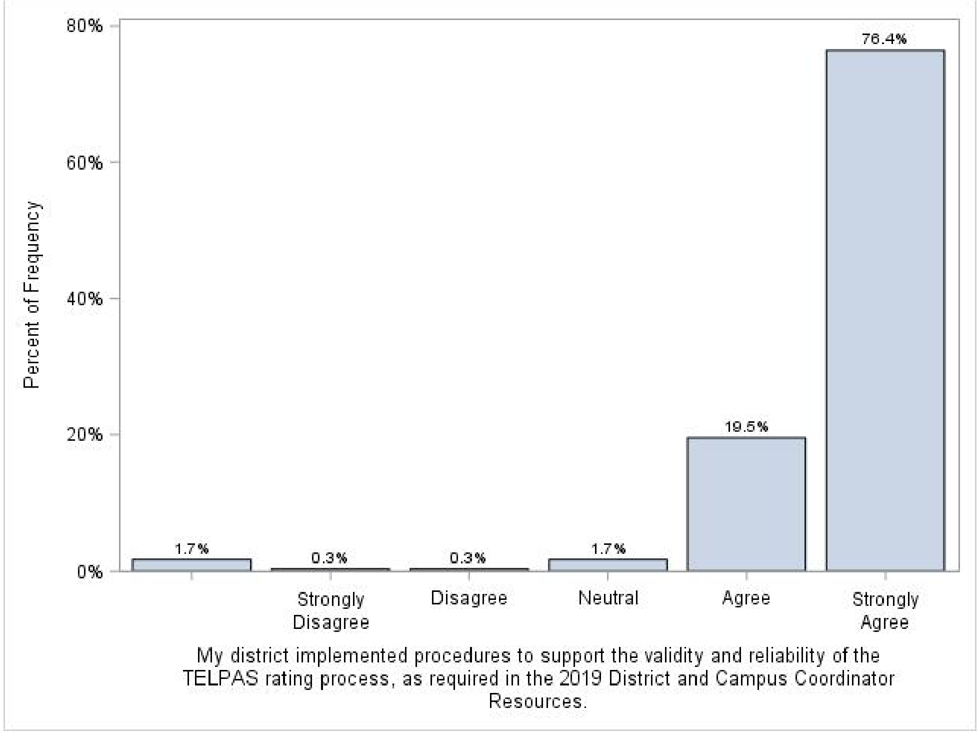
3. Approximately how many individuals served as raters in your district in spring 2019? (Include both new and returning raters.)



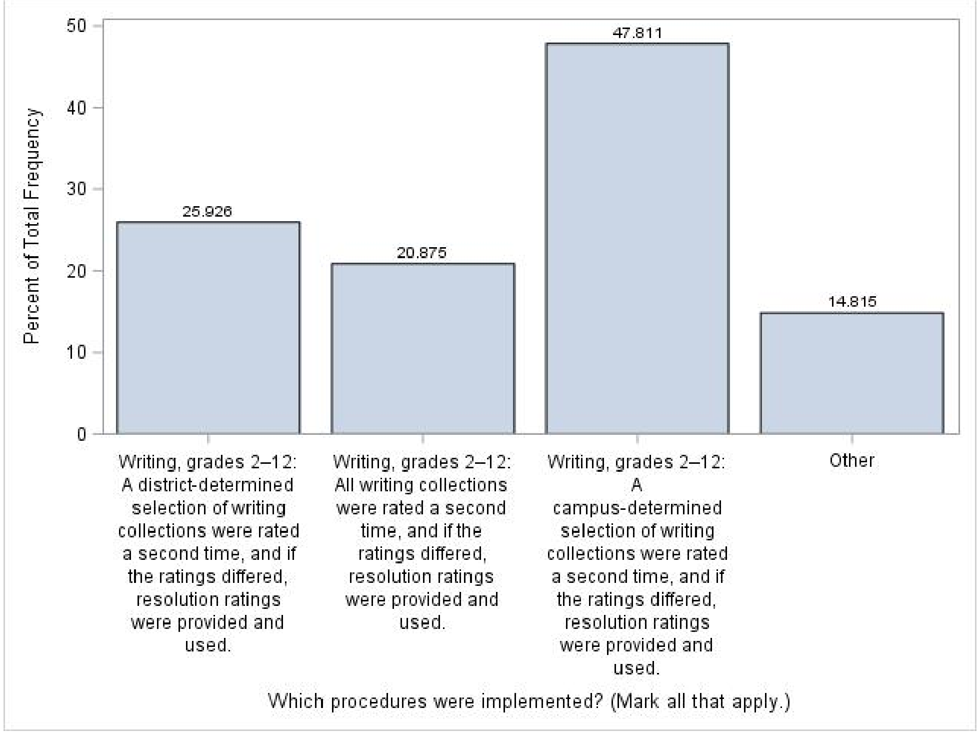
4. All raters in my district who did not successfully calibrate by the end of set 2 were given rating support.



5. My district implemented procedures to support the validity and reliability of the TELPAS rating process, as required in the 2019 District and Campus Coordinator Resources.

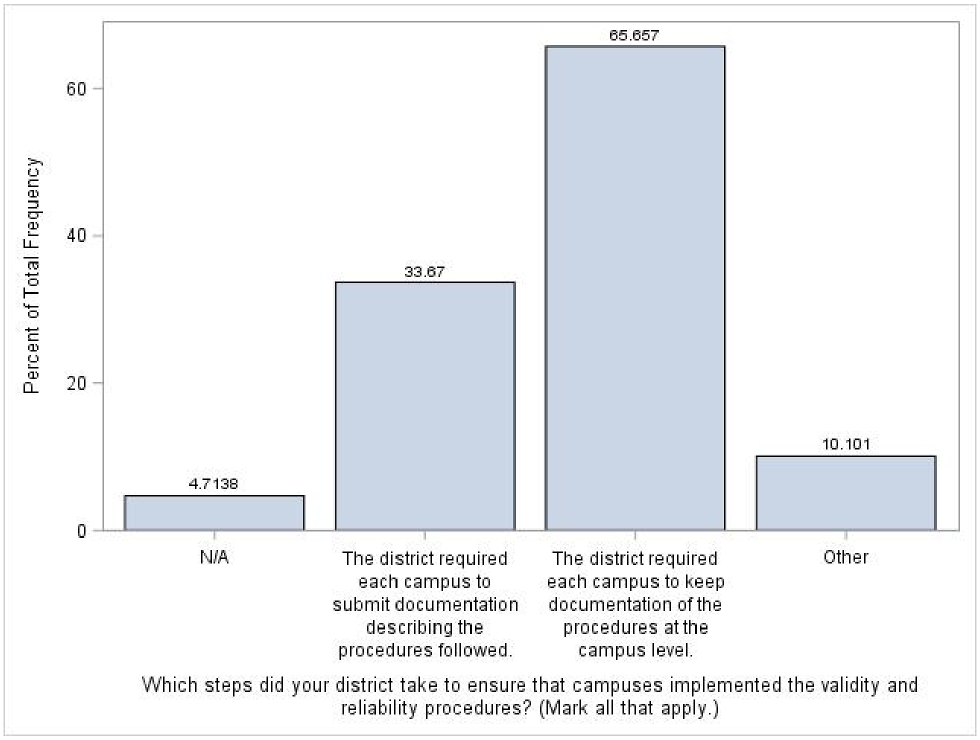


6. What procedures were implemented? (Mark all that apply.)



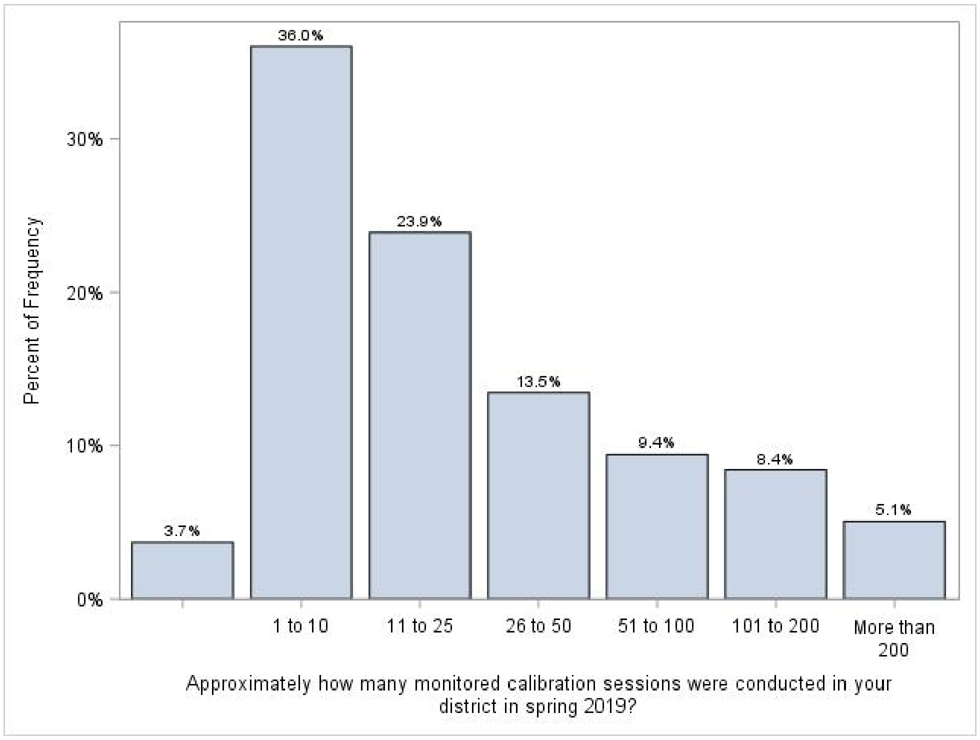
Note: Respondents were allowed to select more than one category. Therefore, totals do not necessarily equal 100%.

8. What steps did your district take to ensure that campuses implemented the validity and reliability procedures? (Mark all that apply.)

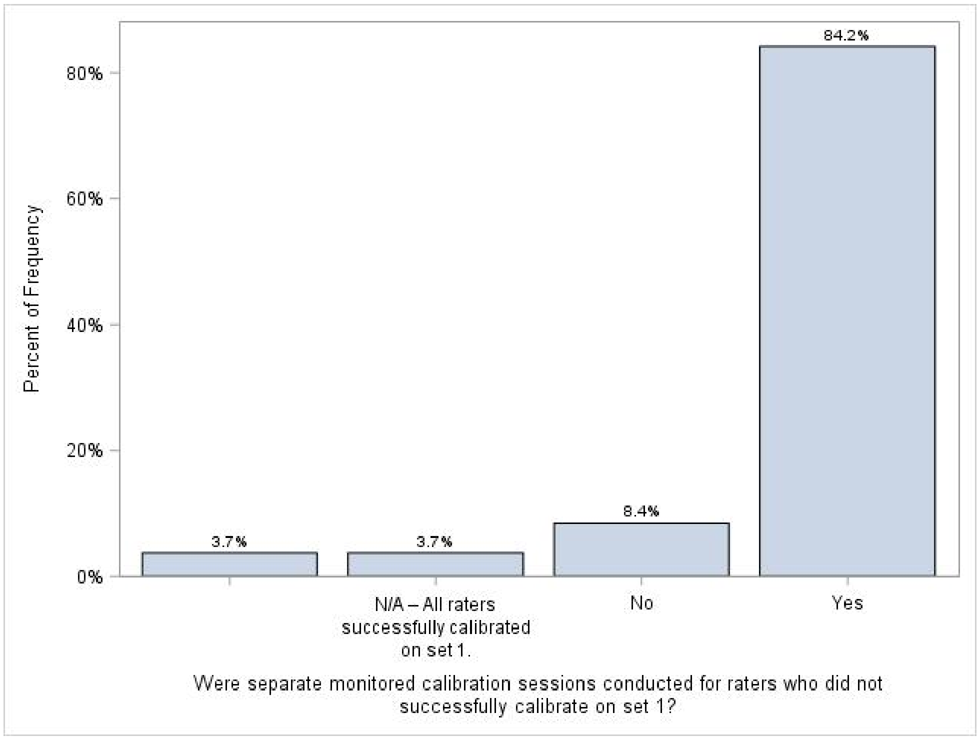


Note: Respondents were allowed to select more than one category. Therefore, totals do not necessarily equal 100%.

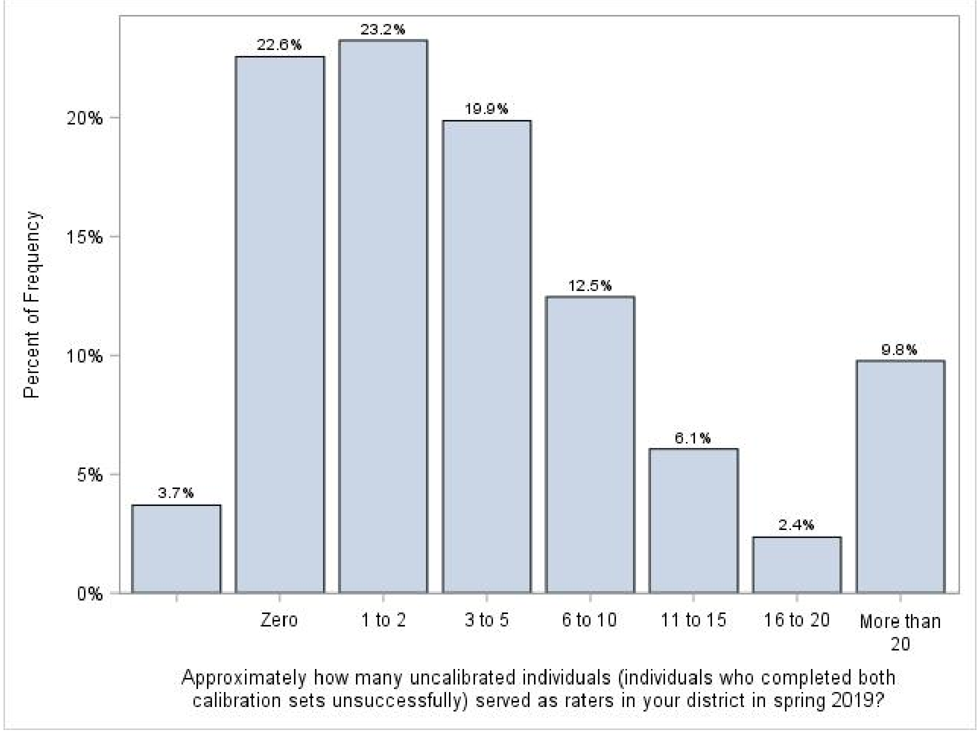
9. Approximately how many monitored calibration sessions were conducted in your district in spring 2019?



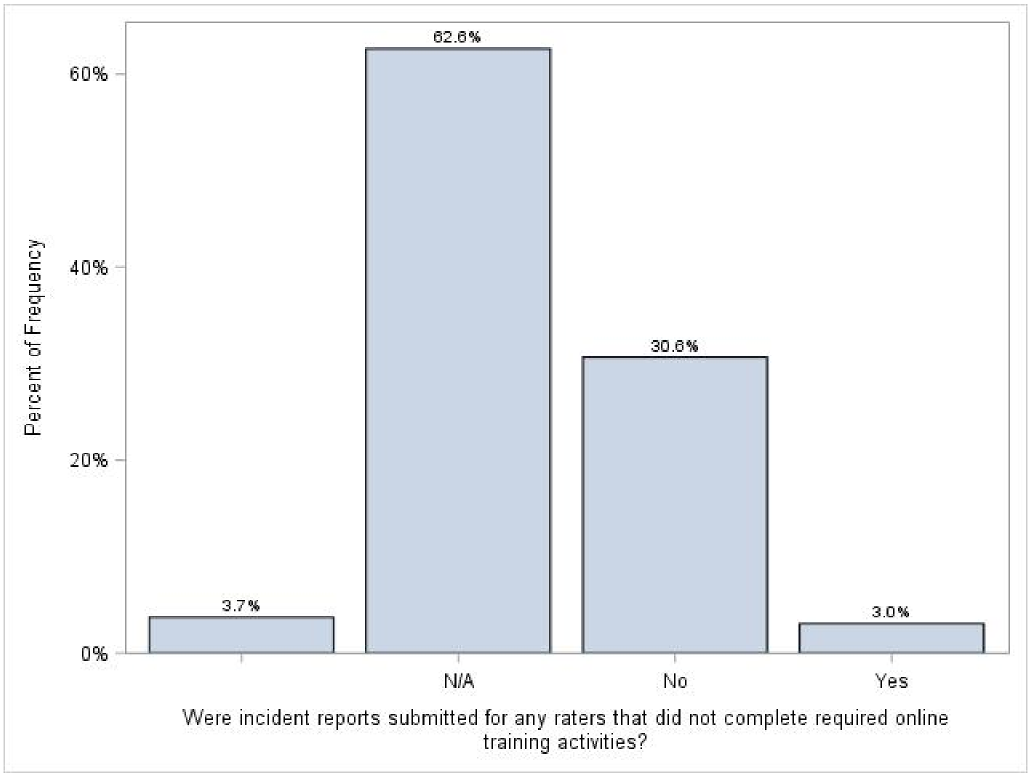
10. Were separate monitored calibration sessions conducted for raters who did not successfully calibrate on set 1?



11. Approximately how many uncalibrated individuals (individuals who completed both calibration sets unsuccessfully) served as raters in your district in spring 2019?



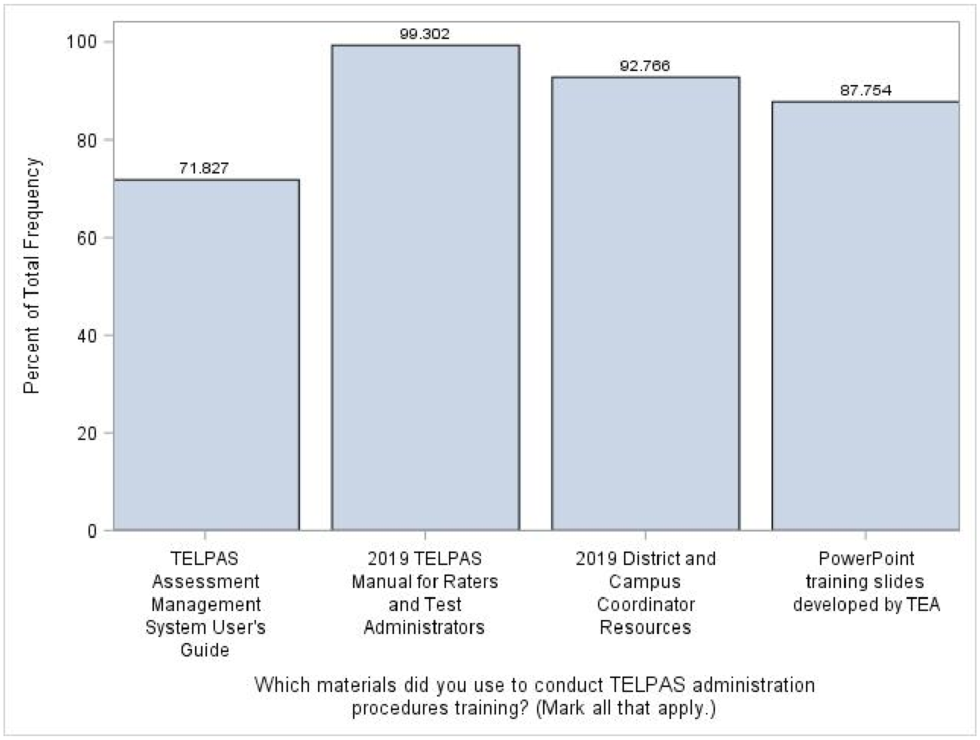
12. Were incident reports submitted for any raters that did not complete required online training activities?



Campus Testing Coordinators

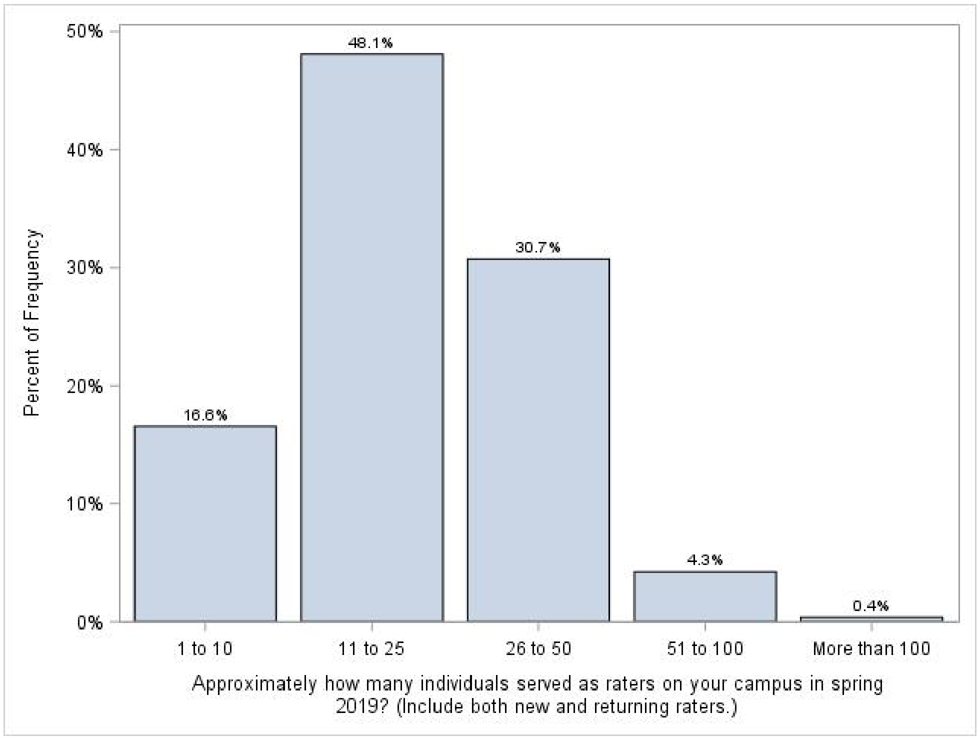
Campus testing coordinators (CTCs) at campuses selected for participation in the 2019 TELPAS writing audit were asked to respond to a set of 13 questions. The first question asked them to select their district and campus names. Questions 2–13 are provided below along with their responses. In general, CTCs reported that proper procedures are in place to train raters and evaluate proper implementation of the holistic ratings. A total of 1,576 responses were provided by CTCs.

2. What materials did you use to conduct TELPAS administration procedures training? (Mark all that apply.)

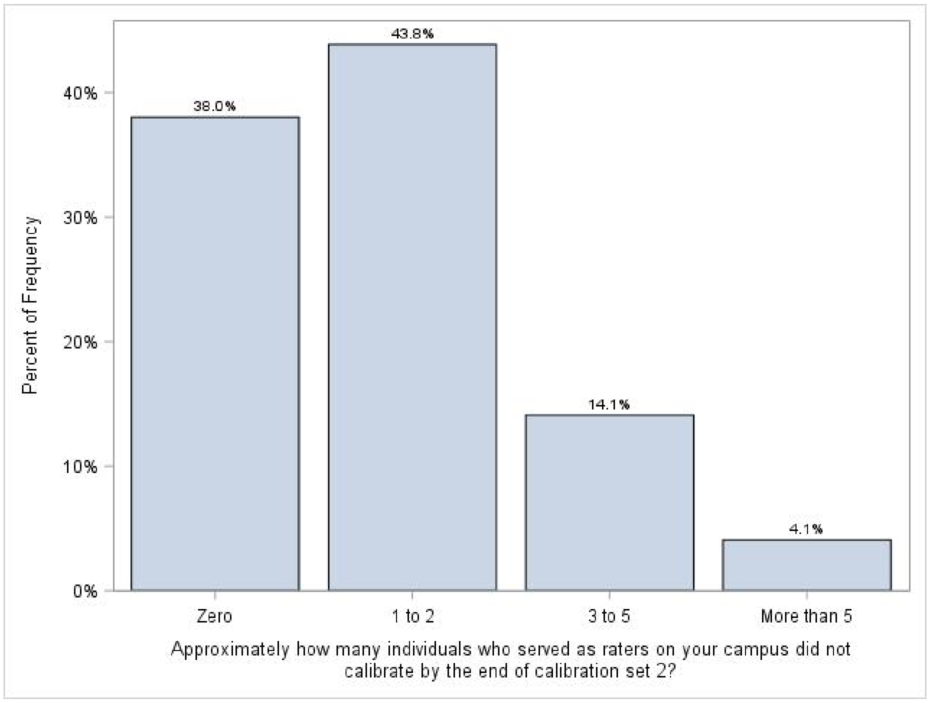


Note: Respondents were allowed to select more than one category. Therefore, totals do not necessarily equal 100%.

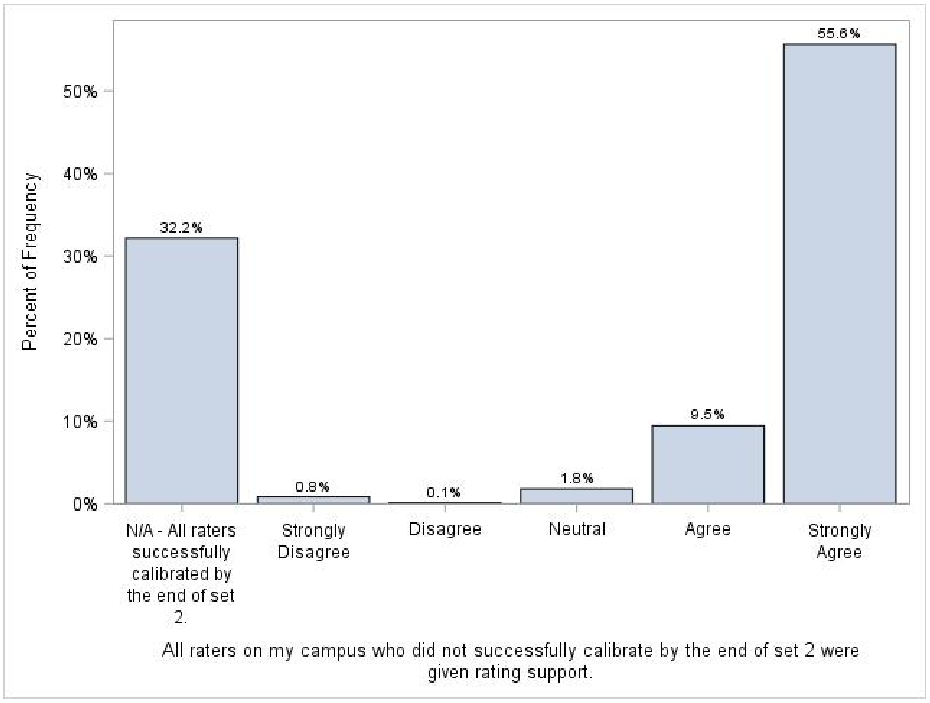
3. Approximately how many individuals served as raters on your campus in spring 2019? (Include both new and returning raters.)



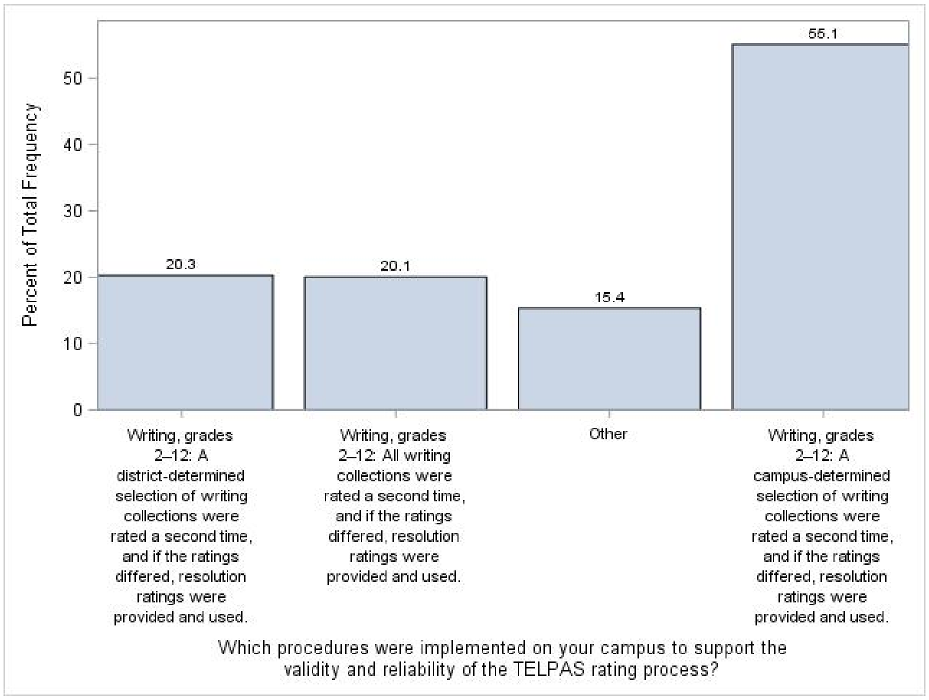
4. Approximately how many individuals who served as raters on your campus did not calibrate by the end of calibration set 2?



5. A raters on my campus who did not successfully calibrate by the end of set 2 were given rating support.

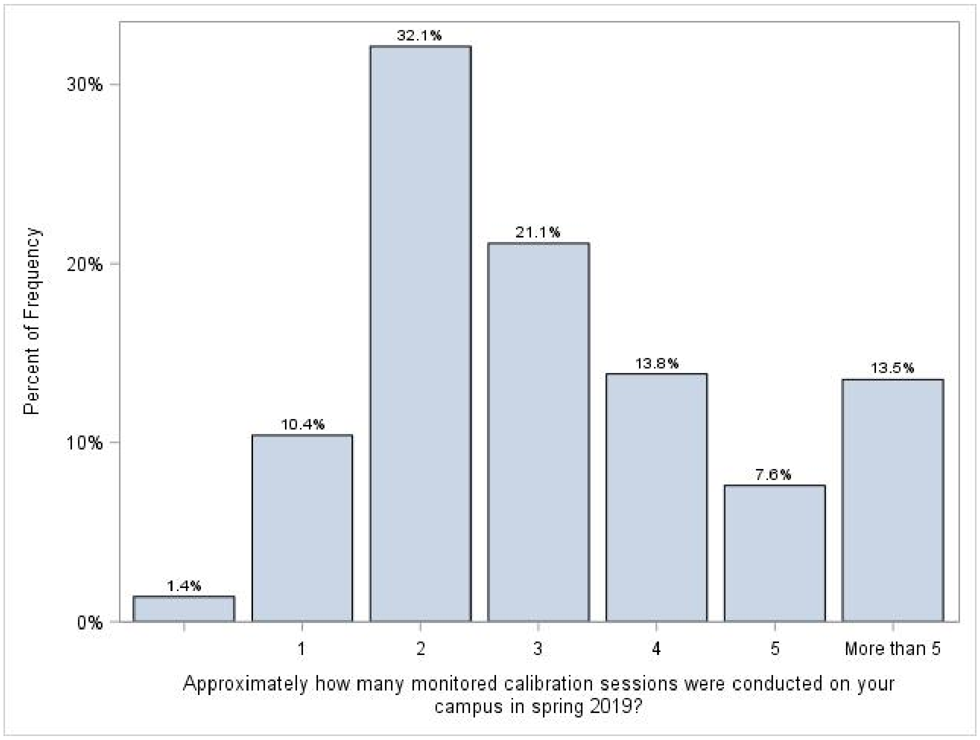


6. Which procedures were implemented on your campus to support the validity and reliability of the TELPAS rating process? Refer to the procedures outlined in the TELPAS Writing Collections section of the 2019 District and Campus Coordinator resources for more information. (Mark all that apply.)

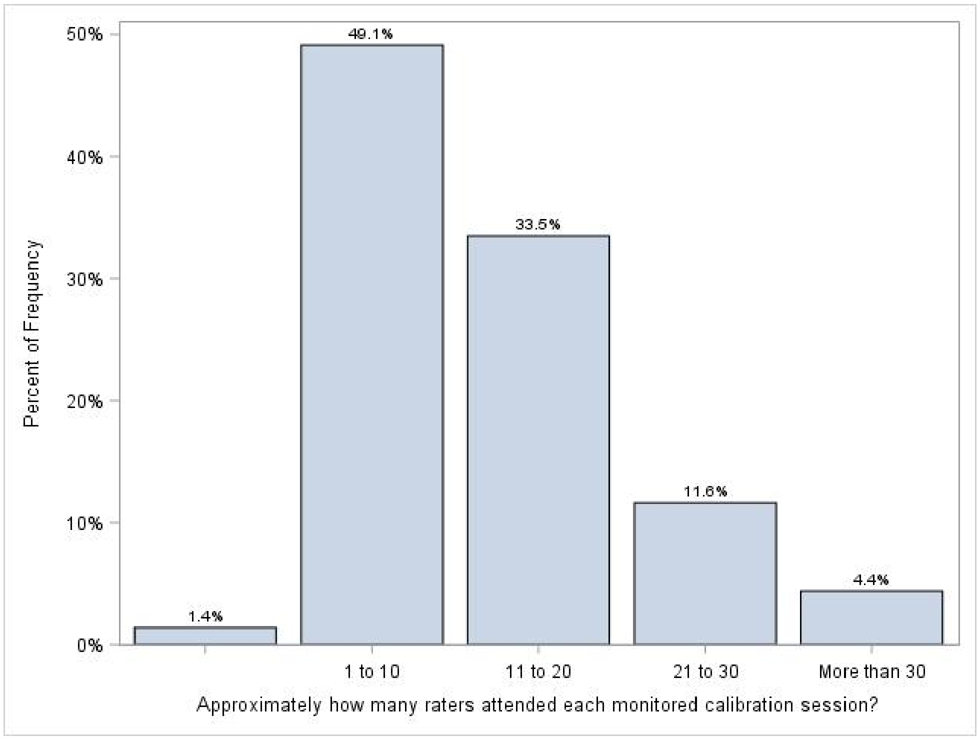


Note: Respondents were allowed to select more than one category. Therefore, totals do not necessarily equal 100%.

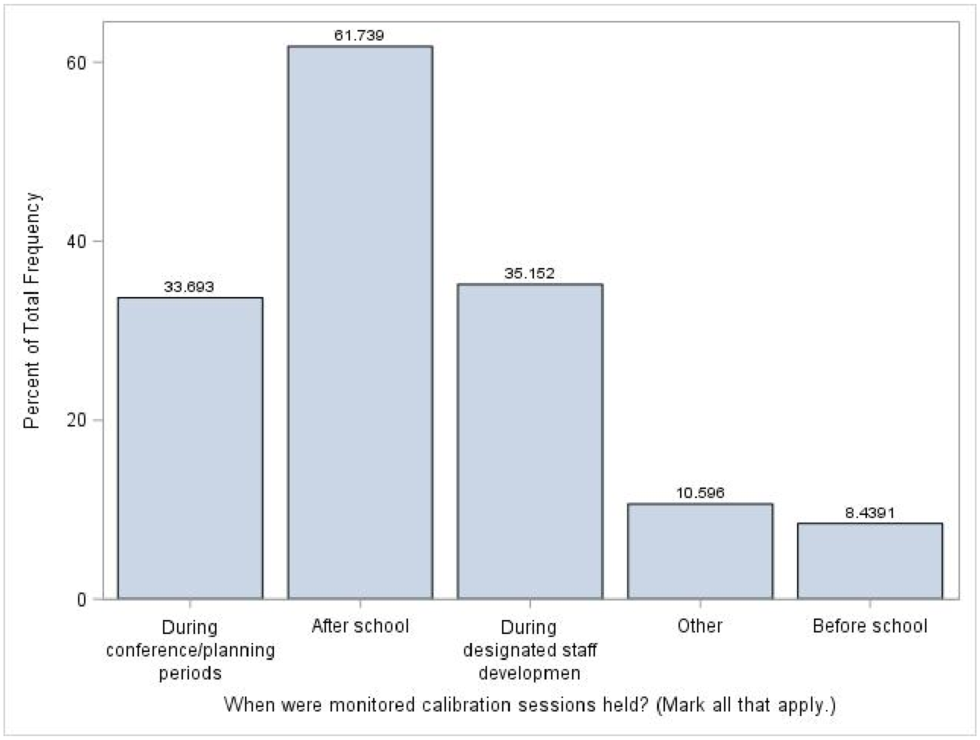
7. Approximately how many monitored calibration sessions were conducted on your campus in spring 2019?



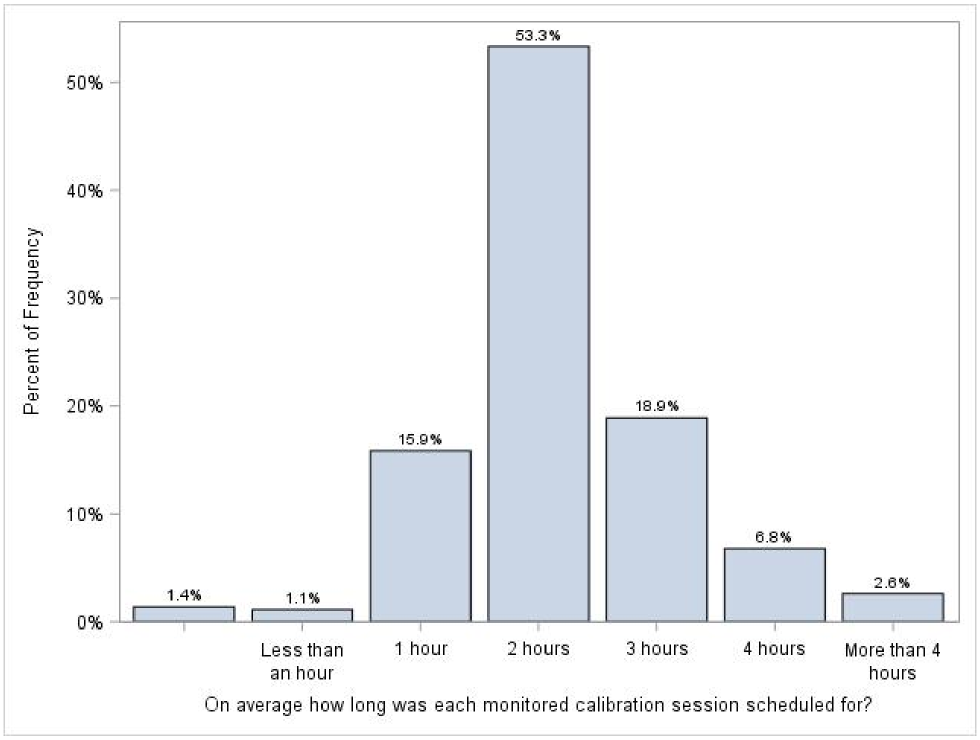
8. Approximately how many raters attended each monitored calibration session?



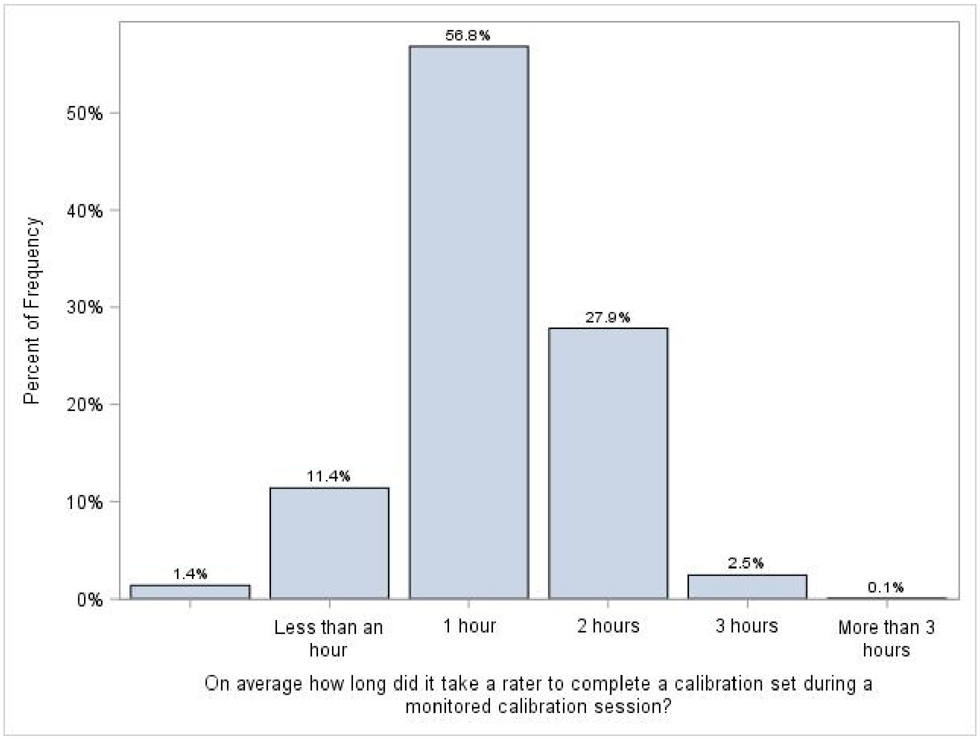
9. When were monitored calibration sessions held?



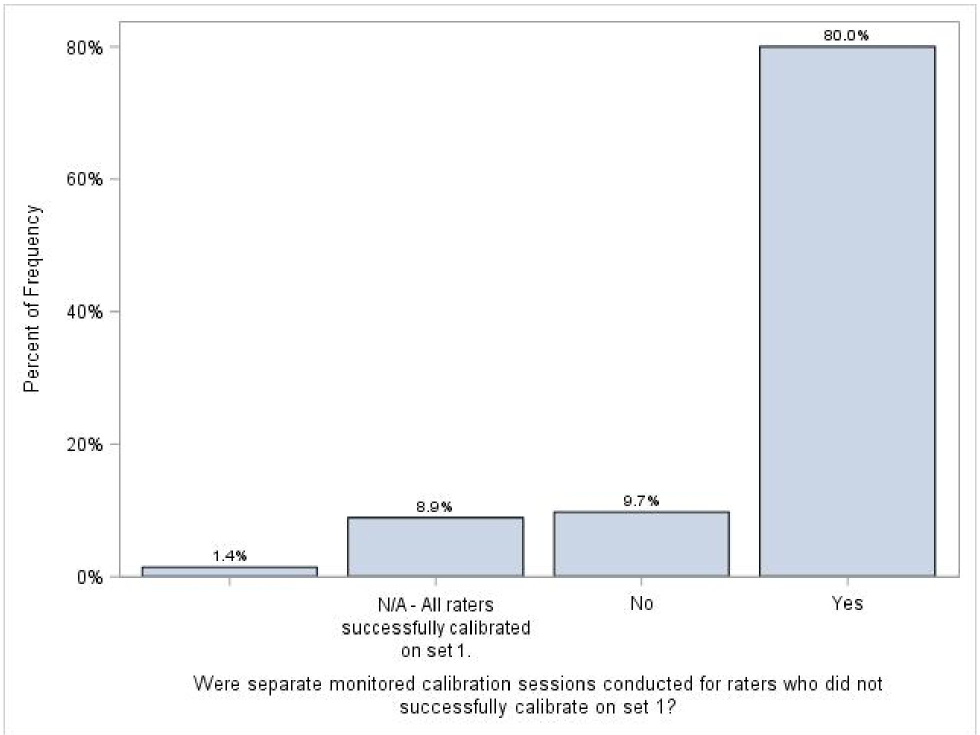
10. On average how long was each monitored calibration session scheduled for?



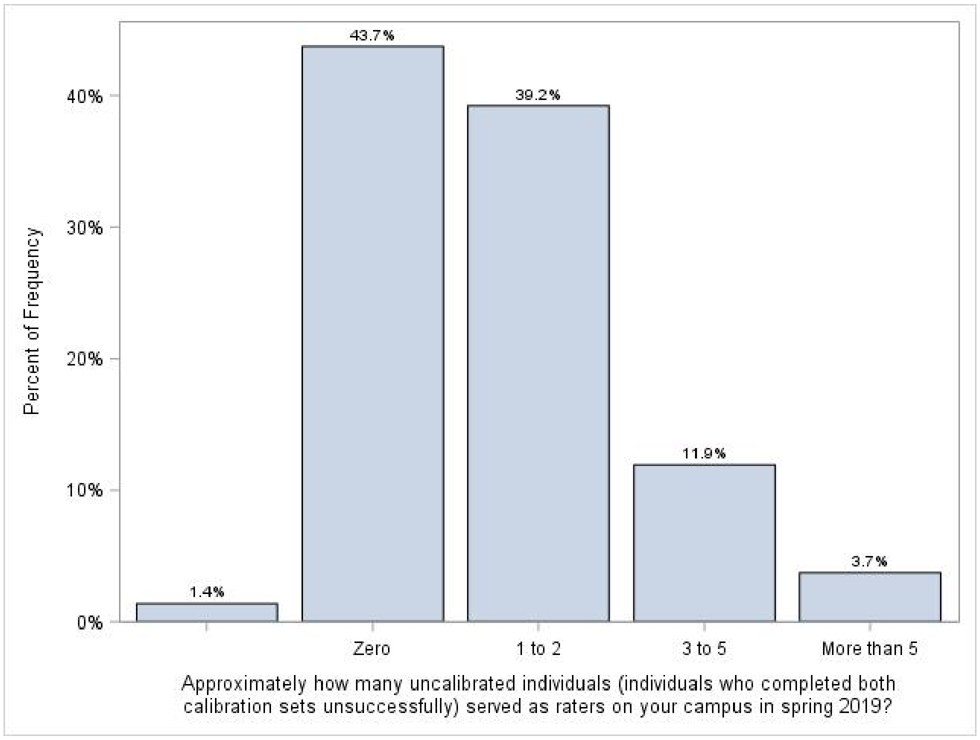
11. On average how long did it take a rater to complete a calibration set during a monitored calibration session?



12. Were separate monitored calibration sessions conducted for raters who did not successfully calibrate on set 1?



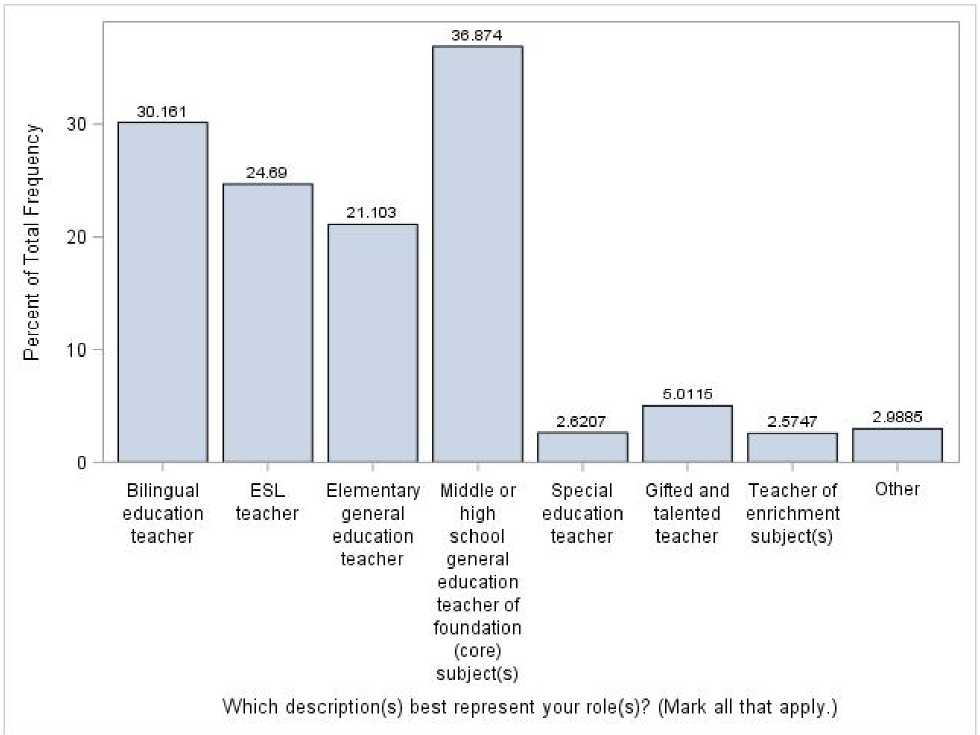
13. Approximately how many uncalibrated individuals (individuals who completed both calibration sets unsuccessfully) served as raters on your campus in spring 2019?



Teacher Raters

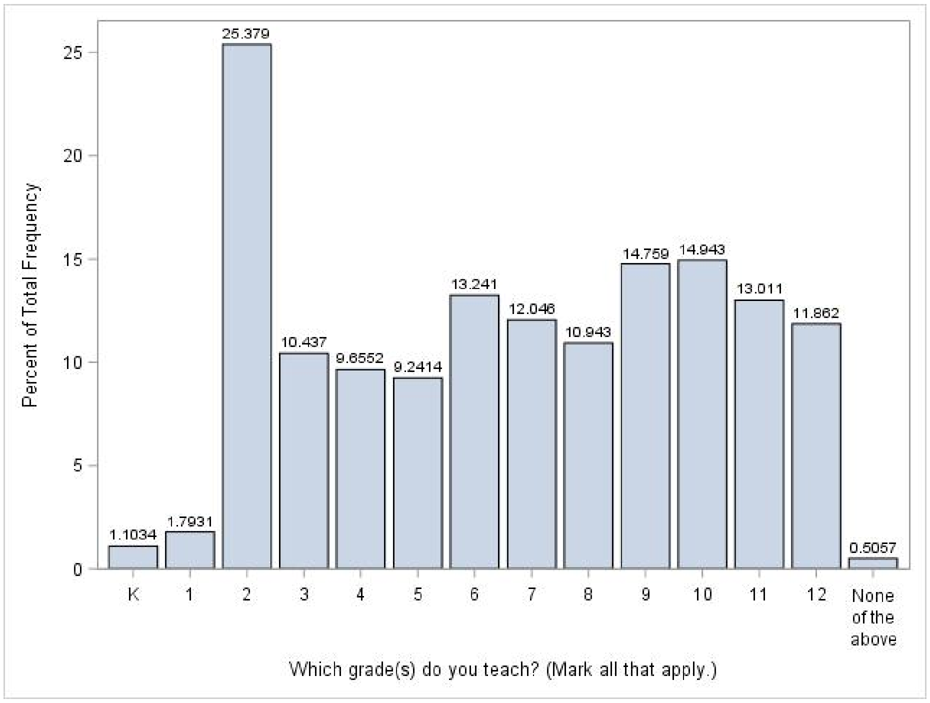
Teacher raters selected for participation in the 2019 TELPAS writing audit were asked to respond to a set of 13 questions. The first question asked them to select their district and campus names. Questions 2–13 are provided below along with their responses. In general, teachers reported receiving adequate training to provide holistic ratings and were confident in the scores they provided. A total of 2,175 responses were provided by teacher raters.

2. Which description(s) best represent your role(s)? (Mark all that apply.)



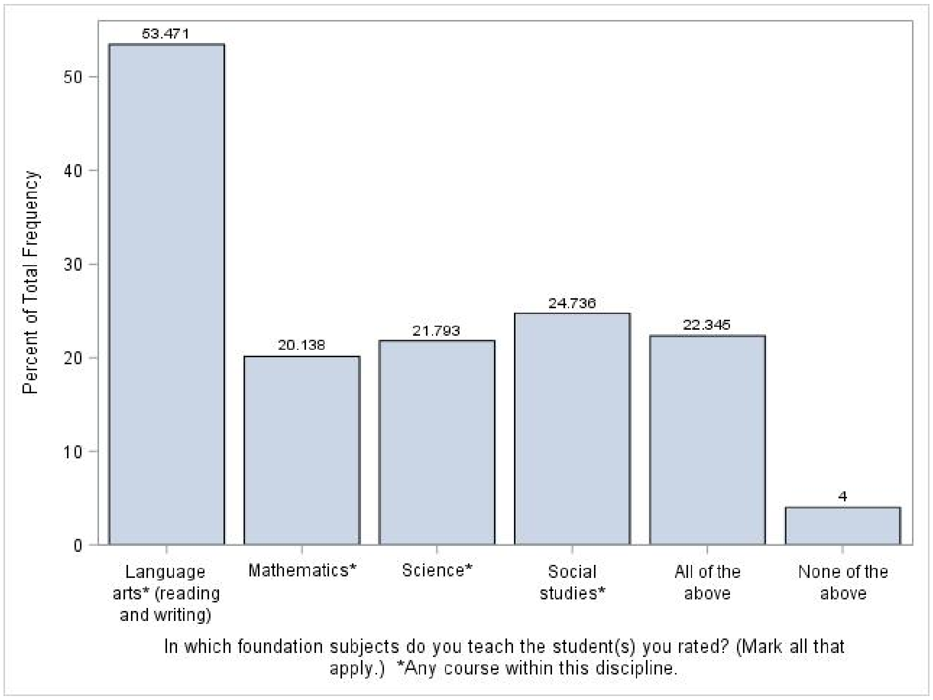
Note: Respondents were allowed to select more than one category. Therefore, totals do not necessarily equal 100%.

3. Which grade(s) do you teach? (Mark all that apply.)



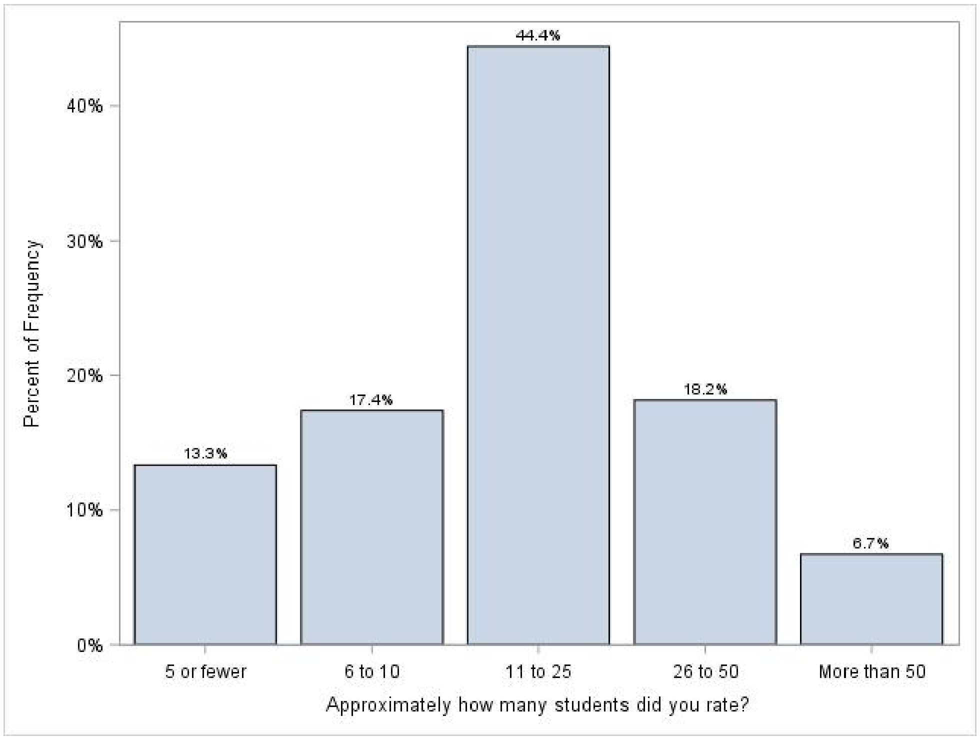
Note: Respondents were allowed to select more than one category. Therefore, totals do not necessarily equal 100%.

4. In which foundation subjects do you teach the student(s) you rated? (Mark all that apply.) *Any course within this discipline.

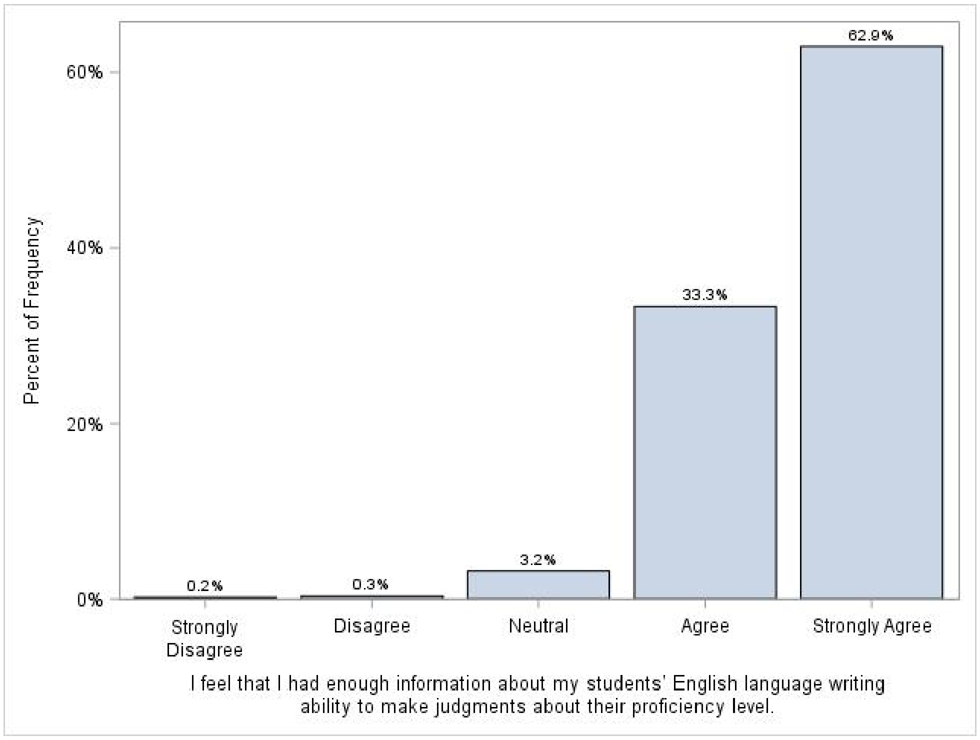


Note: Respondents were allowed to select more than one category. Therefore, totals do not necessarily equal 100%.

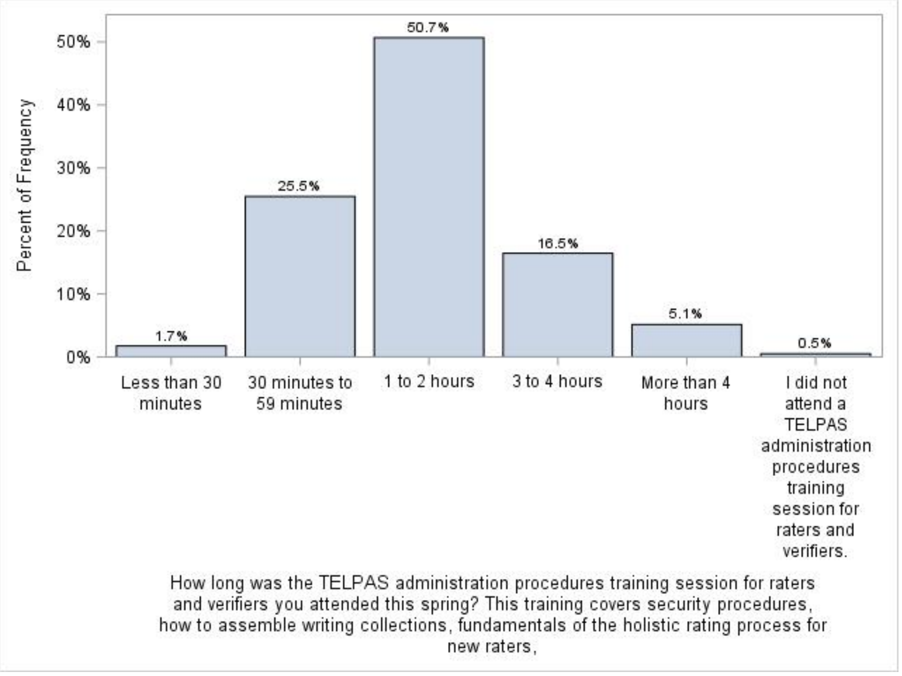
5. Approximately how many students did you rate?



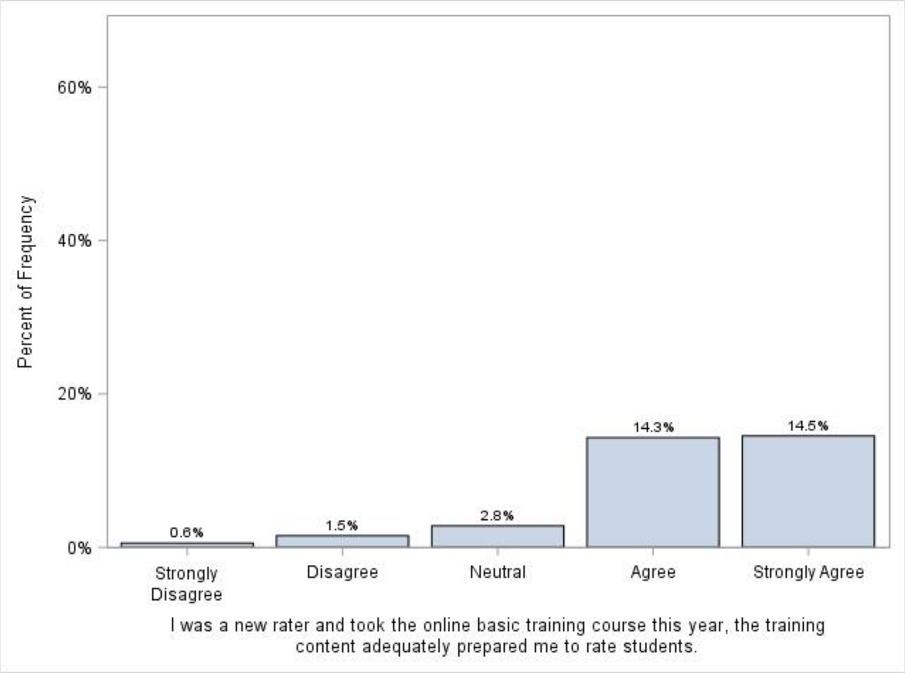
6. I feel that I had enough information about my students' English language writing ability to make judgments about their proficiency level.



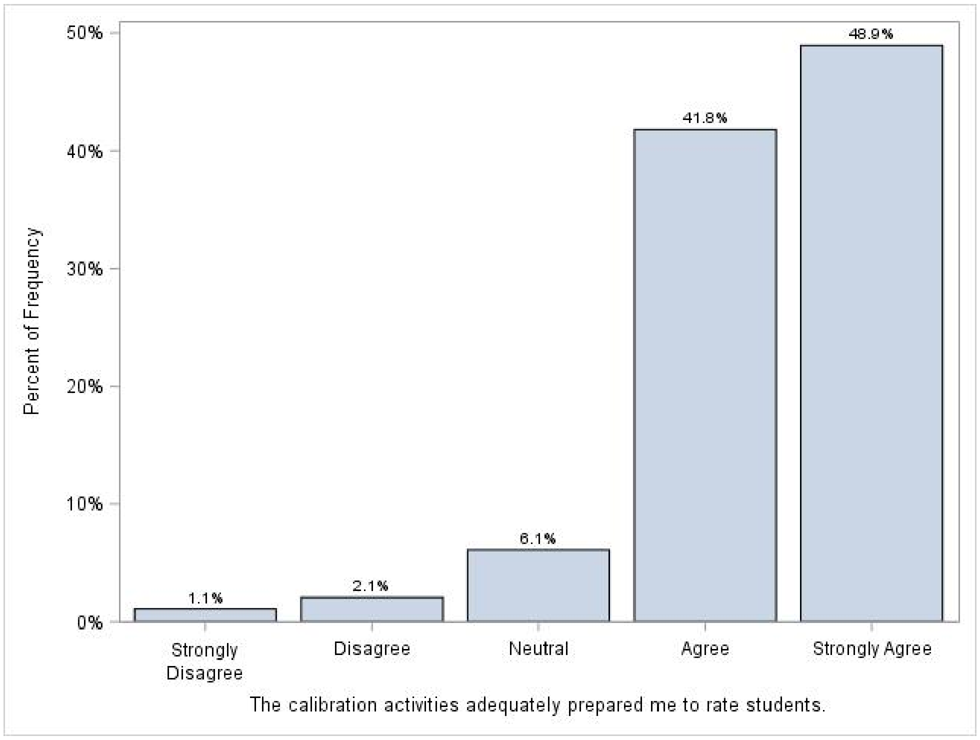
7. How long was the TELPAS administration procedures training session for raters and verifiers you attended this spring? This training covers security procedures, how to assemble writing collections, fundamentals of the holistic rating process for new raters, holistic rating training requirements for raters, etc. If this training was conducted in conjunction with other training, please indicate just the approximate amount of time spent on TELPAS administration procedures. If you completed the Assembling and Verifying Grades 2-12 Writing Collections online course, do not include the completion time for that course in your response.



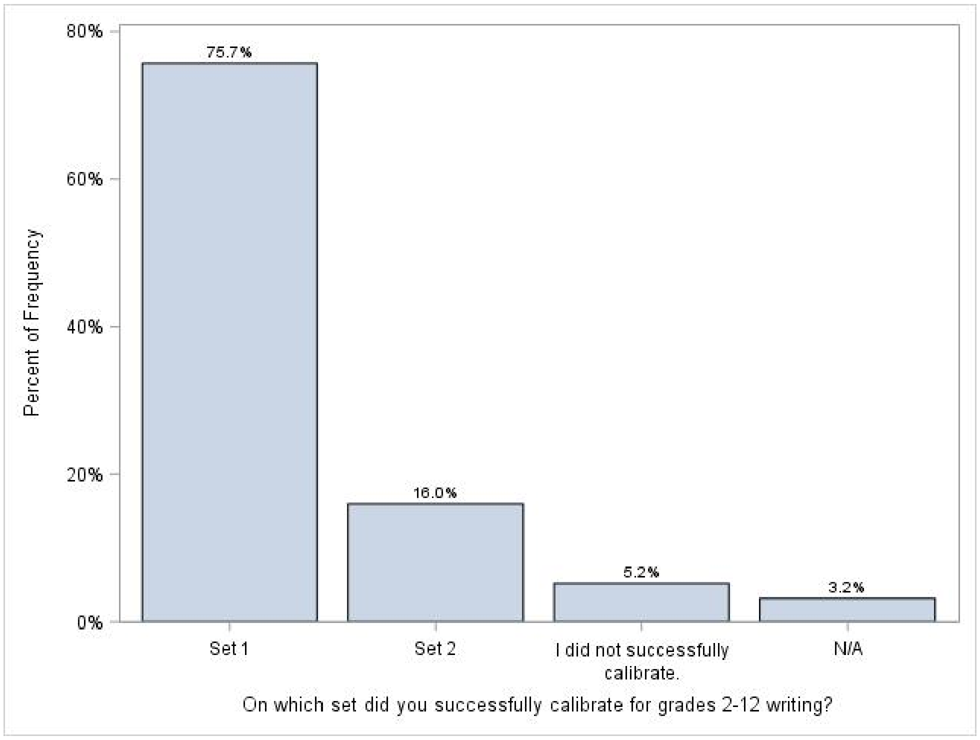
8. I was a new rater and took the online basic training course this year, the training content adequately prepared me to rate students.



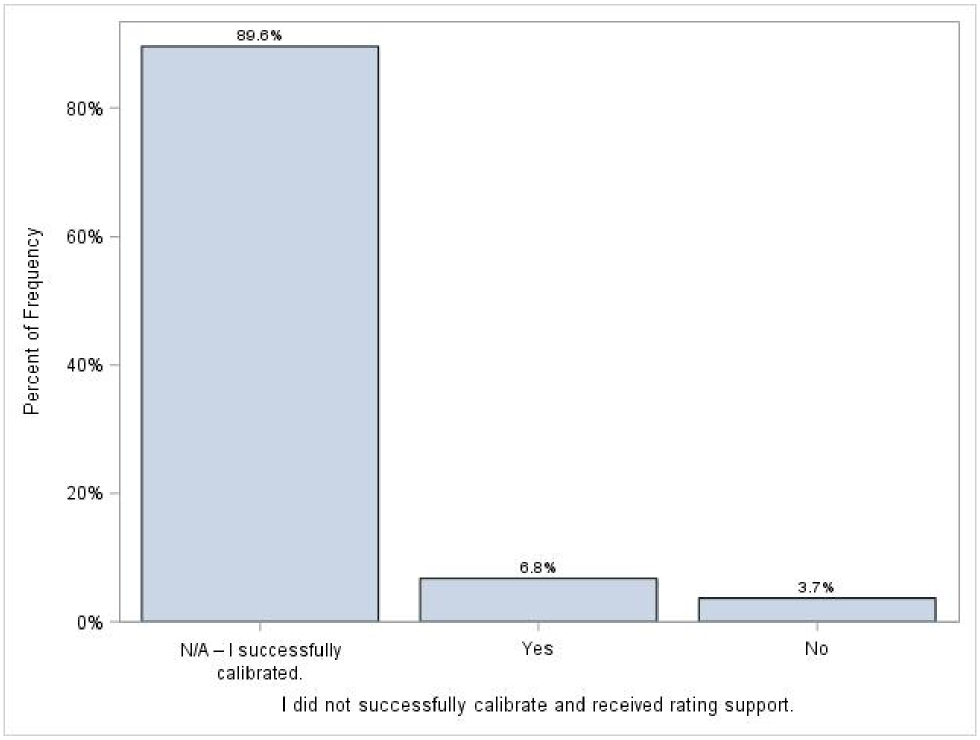
9. The calibration activities adequately prepared me to rate students.



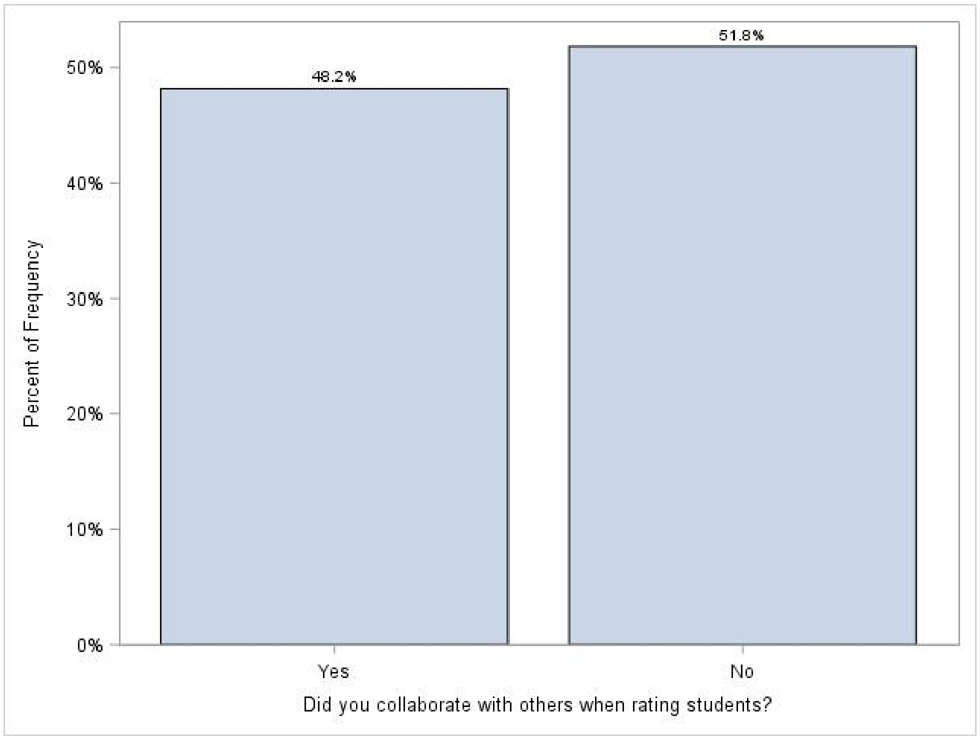
10. On which set did you successfully calibrate for grades 2-12 writing?



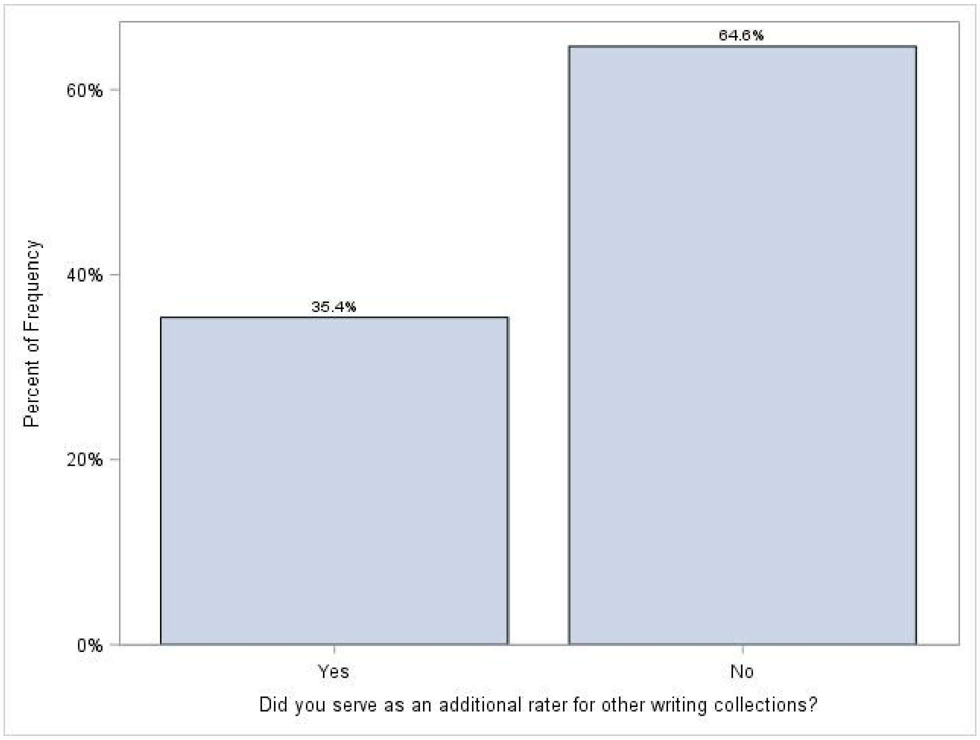
11. I did not successfully calibrate and received rating support.



12. Did you collaborate with others when rating students?



13. Did you serve as an additional rater for other writing collections?


Percent of Frequency
Yes
No
Did you serve as an additional rater for other writing collections?