June 2019
Publication 18.17

Measurement Validity and Reliability of Professional Pathways for Teachers: Technical Report

PPfT

Department of
Research and Evaluation
AUSTIN INDEPENDENT SCHOOL DISTRICT

Executive Summary

The purpose of this Professional Pathways for Teachers (PPfT) evaluation was to examine the measurement validity and reliability of PPfT appraisal data from the 2017–2018 school year. The PPfT appraisal is a multi-measure system that covers three areas: instructional practices (IP), professional growth and responsibilities (PGR), and two student growth measures: a teacher-level student learning objective (SLO) measure and a campus-level school-wide value-added (SWVA) measure.

The validity and reliability of the PPfT appraisal related to two basic ideas: did we measure what we intended to measure and can we measure it consistently? To address the validity question we examined content validity, concurrent validity, convergent validity, discriminant validity, and dominance. To address the reliability question we examined interrater reliability and internal consistency.

Evidence suggested strong content validity around the entire instructional practices process. However, stakeholders seemed divided on their perceptions of whether the appraisal system measures teaching quality. Issues around item design and education or uncertainty were considered.

The overall differentiation of teachers across final rating categories (i.e., distinguished, highly effective, effective, minimally effective, and ineffective) suggests inter-category concurrent validity. However, the large mode of teachers receiving a highly effective final rating suggests weaker intra-category concurrent validity. A shift in the procedures for rater calibration around differentiating 3s and 4s on the instructional practice rubric to what 2s and 3s look like in the classroom was explored as means to maintain inter-category concurrently validity while potentially improving intra-category concurrent validity.

Evidence suggested strong convergent validity of final ratings. For most grades and subjects examined, correlation analyses showed that as teaching quality increased, so did student growth.

Evidence suggested mixed discriminant validity findings across the student characteristics observed. The gender of the students served by teachers, gifted and talented (GT) status, and special education (SPED) status appeared to operate independently of the final ratings teachers received. However, the limited English proficiency (LEP) status, economically disadvantaged (ECONDIS) status, and the race/ethnicity of the students served by teachers appeared to operate in some dependency with the final ratings teachers received. The strategic recruiting and compensation of the Comprehensive Schools Improvement Model was explored as a potential lever to equitably distribute high-quality teachers with populations of underserved students.

Dominance analysis revealed that IP ratings were the most important contributor to predicting final ratings, followed by SLO ratings, PGR ratings, and lastly SWVA ratings. Analyses suggest that very little additional information is being added by PGR and IP ratings over IP ratings alone. Results further underscored the importance of distribution quality (e.g., variance and normality) for each PPfT appraisal component.

Interrater reliability analyses were inconclusive due to confounds between raters and time and teacher improvement. However, collectively, the set analyses of interrater reliability were interpreted to suggest adequate agreement between raters. Use of floating peer observers to partner with school administrators during both observations was explored as means to reduce confounds in analysis of interrater reliability.

Fall IP ratings, spring IP ratings, and PGR ratings showed evidence of strong internal consistency, but the set of four appraisal components (i.e., IP, PGR, SLOs, and SWVA) showed evidence of somewhat weak internal consistency. Although internal consistency did not meaningfully improve with removal of any components, exploratory analysis considering replacement of the SWVA component with a teacher value-added component did meaningfully improve internal consistency and change the factor analytic structure to a single factor solution.

Table of Contents

Introduction

Purpose of Evaluation

This technical report is a supplement to DRE Publication 18.17 RB (Hutchins, 2019) and a follow-up study to DRE Publication 17.60 (Hutchins, Looby, DeBaylo, & Leung, 2019). The purpose of this Professional Pathways for Teachers (PPfT) evaluation was to examine the measurement validity and reliability of PPfT appraisal data from the 2017–2018 school year. The paper was prepared in response to questions from the PPfT oversight committee, district leadership, and program staff.

Description of PPfT

PPfT is a human capital system that blends four primary components: appraisal, professional development (PD) opportunities, leadership opportunities, and compensation. The goal of PPfT is to build the capacity of Austin Independent School District (AISD) teachers through a comprehensive system of supports and compensation. Underlying this goal is the core belief that professionalizing teaching and empowering teachers through comprehensive supports and compensation will lead to positive impacts on teacher retention and student achievement.

Description of Validity and Reliability Study

The 2017–2018 appraisal component of PPfT was the subject matter of this measurement validity and reliability study. The PPfT appraisal is a multi-measure system with four primary components that covers three areas: instructional practices (IP), professional growth and responsibilities (PGR), and student growth. Student growth includes two measures: a teacher-level student learning objective (SLO) measure and a campus-level school-wide value-added (SWVA) measure. The validity and reliability of the PPfT appraisal relates to two basic ideas; that is, did we measure what we intended to measure, and can we measure it consistently? Several specific questions framed the analysis:

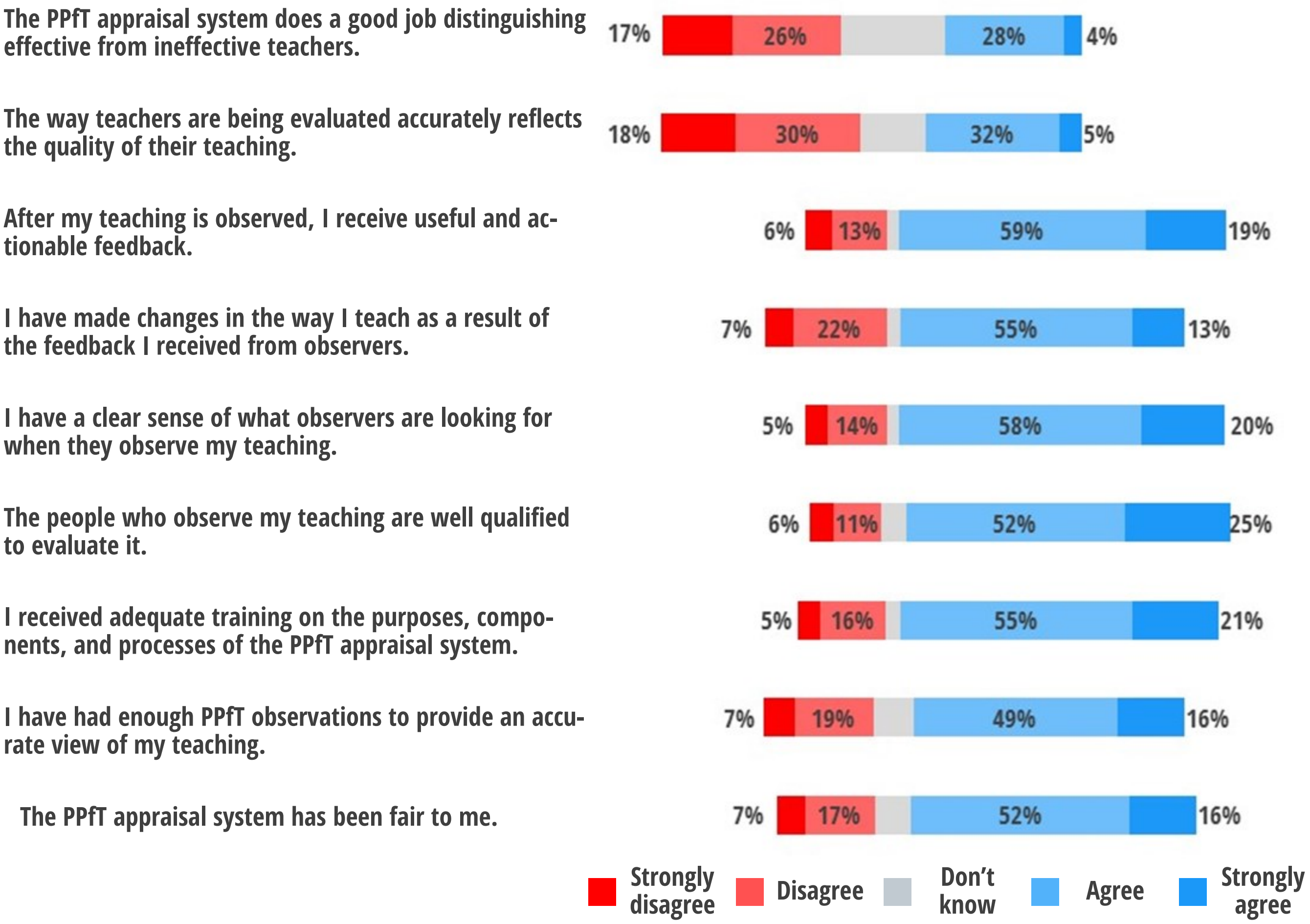
> Did stakeholders feel the final and IP ratings reflected the quality of their teaching (i.e., content validity)?
>
> To what extent did PPfT final ratings and individual appraisal components differentiate teachers (i.e., concurrent validity)?
>
> To what extent were teachers' PPfT final ratings PPfT associated with their students' growth (i.e., convergent validity)?
>
> To what extent were teachers' PPfT final ratings associated with their students' characteristics (i.e., discriminant validity)?
>
> What components of PPfT appraisal were most important to prediction of PPfT final ratings (i.e., dominance)?
>
> Did ratings of the same teacher vary between different raters (i.e., inter-rater reliability)?
>
> To what extent were strand ratings within components correlated and to what extent were the components ratings of PPfT correlated (i.e., internal consistency)?

Validity and Reliability Results

Content Validity: Did stakeholders feel the final and instructional practice ratings reflected the quality of their teaching?

Survey responses to numerous items collected from teachers in the spring of 2018 provided evidence for teachers' perceptions of the content validity of the PPfT appraisal and IP ratings. Two patterns emerge when looking at the spread of negative and positive responses away from the neutral response (Figure 1). First, most teachers felt there was legitimacy to their IP ratings and that the appraisal system was fair to them. Second, teachers were divided on whether the appraisal system distinguished teacher effectiveness and was an accurate reflection of teaching quality.

Figure 1.
Most teachers felt there was legitimacy to their IP ratings, but teachers were divided on whether the appraisal system reflected teacher effectiveness and teaching quality.



Source. 2017–2018 Employee Coordinated Survey.
Note. Of the 5,577 teachers appraised under PPfT in 2017–2018, approximately 8% (*n* = 470) responded to the survey.

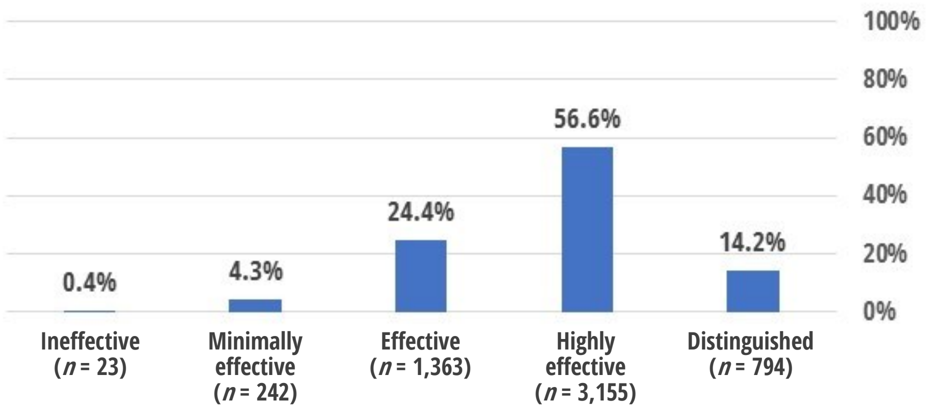
Survey results suggest strong content validity for IP observations. However, survey results are less conclusive for the overall appraisal system. A few interpretations are offered. Differences in how teachers responded to the fairness statement and the teacher effectiveness and teaching quality statements suggest that teachers think of fairness and whether the appraisal system measures teaching quality as two different ideas. On the two items where teachers were divided on their perceptions of validity (i.e., teacher effectiveness and teaching quality), the items are broad third-person statements about all teachers as opposed to the first-person (i.e., *I* and *me*) IP statements. The difference in response patterns suggests that teachers respond differently depending on item design. Furthermore, there was a spike in don't know responses, up to as much as 25%, on these two items. This suggests that the mixed perceptions were an issue of education and uncertainty rather than a firm division between teachers.

Concurrent Validity: To what extent did final ratings on PPfT differentiate teachers?

In a more traditional sense, concurrent validity would assess whether some semi-parallel instrument that measures quality of teaching differentiates teaching quality with approximately the same groupings of teachers as does the PPfT appraisal (e.g., higher-quality teaching on the Texas Teacher Evaluation and Support System (T-TESS) would be found for similar groups of teachers on the PPfT appraisal and vice versa for lower-quality teaching). Absent another existing instrument of teaching quality on which AISD teachers were measured, the construct of concurrent validity more simply assessed whether the existing PPfT instrument differentiated teachers. In this alternative sense, concurrently validity would not be demonstrated if all teachers were distinguished or otherwise received very similar final ratings.

To assess the concurrent validity of final ratings, the distribution of 2017–2018 PPfT final ratings was examined (Figure 2). The distribution of PPfT final ratings shows a clear differentiation of teachers across all final rating categories. However, 56.6% of all teachers appraised in 2017–2018 received a highly effective final rating. To assess the concurrent validity of individual PPfT appraisal component scores, the distributions of 2017–2018 ratings were examined (Figure 3). Differentiation between teachers at the component-level was best for SWVA and SLOs, worst for IP and PGR.

Figure 2.
PPfT final ratings differentiated teachers, but appraised more than half of teachers (i.e., 56.6%) as highly effective.



Source. 2017–2018 Employee Coordinated Survey.

PPfT Versus T-TESS

Differences between PPfT and T-TESS can be confusing. When comparing the two systems, it must be kept in mind that PPfT is an entire human capital system inclusive of four primary components: teacher appraisal, teacher PD opportunities, teacher leadership opportunities, and teacher compensation. T-TESS is the teacher evaluation system adopted by the Texas Education Agency (TEA). PPfT meets and exceeds the state requirements for teacher appraisal.

Given the difference in scope of PPfT and T-TESS, comparison requires limiting PPfT to only the teacher appraisal component.

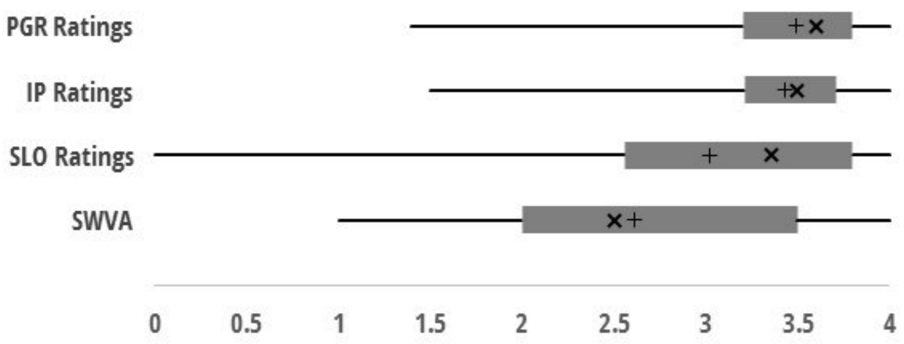
T-TESS includes three primary components: a goal-setting and professional development plan, an evaluation cycle, and student growth. T-TESS results in one of five performance level ratings for teachers:

- **Distinguished**
- **Accomplished**
- **Proficient**
- **Developing**
- **Improvement Needed**

The PPfT appraisal component of the overall PPfT human capital system includes four primary components: instructional practices, professional growth and responsibilities, student learning objective, and school-wide value-added. PPfT results in one of five final ratings for teachers:

- **Distinguished**
- **Highly effective**
- **Effective**
- **Minimally effective**
- **Ineffective**

Figure 3.
Among the components of PPfT final ratings, differentiation was best for SWVA and SLOs, worst for IP and PGR differentiated teachers.



Source. 2017–2018 Employee Coordinated Survey.
Note. Interquartile range, where, X = median, + = mean. PPfT final ratings, IP, and PGR included all 5,577 teachers appraised in 2017–2018. SLO scores included the 5,413 teachers on a new teacher or standard PPfT appraisal plan. SWVA scores included the 4,515 teachers on the standard PPfT appraisal plan.

Convergent Validity: To what extent were teachers' final ratings on PPfT associated with their students' growth?

In this study, it was assumed that quality teaching is associated with student academic growth. Therefore, if PPfT final ratings measured quality teaching, then they should correlate with student academic growth. Convergent validity was examined with the subset of teachers with students testing on the State of Texas Assessments of Academic Readiness (STAAR) grades 3 through 8 and end-of-course (EOCs) assessments. Assessments results were used to model estimates of teacher-level student growth in grades 4 through 8 and on EOCs.

Overall, results of correlation analyses generally showed positive associations between teachers' PPfT final ratings and the academic growth of the teachers' students. Thus, higher-quality teaching (as measured by PPfT) was associated with greater student growth than was lower-quality teaching. More specifically, when comparing final ratings and student growth, convergent validity was observed in grades 4 through 7 in all tested subjects and on the algebra I, biology, English I, and English II EOCs. Convergent validity was generally not observed in grade 8 and on the U.S. history EOC (Table 1, see also Appendix A).

Significant, positive correlations between PPfT final ratings and student growth in both math and reading were observed in grades 4 through 7.

Differentiating Teachers

PPfT Final Ratings

The large group of teachers with highly effective ratings (~57%) calls into question whether these teachers are adequately differentiated by the highly effective category. To think of this differently, if we were to review these 3,155 highly effective teachers, would an informed judge consider them all equivalent in teaching quality?

The overall differentiation of teachers across final rating categories suggests inter-category concurrent validity. However, the large mode of teachers receiving a highly effective final rating suggests weaker intra-category concurrent validity for the highly effective final rating.

PPfT Appraisal Components

At the component-level, differences in differentiation are observed by examining the overall width of each distribution and the relative location of the middle score. IP showed the weakest component-level concurrent validity. IP used a 1 to 4 scale; however, the lowest score earned by any teacher on IP was a 1.5, the highest was a 4, so 2.5 IP points differentiated all 5,577 teachers appraised under PPfT in 2017–2018 in their IP. Given the median of 3.5, only 0.5 of a point differentiated the IP scores of the upper 50% of teachers, and 2 points differentiated the bottom 50%.

In contrast, SWVA showed evidence for the strongest component-level concurrent validity. SWVA used a 4-point scale from 1 to 4, the minimum score earned was 1, the maximum earned was 4, and the median was 2.5, so 1.5 points equally differentiated the upper and lower 50% of teachers' SWVA scores.

Table 1.
In general, higher-quality teaching was associated with greater student growth than was lower-quality teaching.

Tested subject	Tested grades					
	4	5	6	7	8	Secondary (EOC)
Math	+	+	+	+	NR	NA
Reading	+	+	+	+	NR	NA
Science	NA	+	NA	NA	NR	NA
Writing	NA	NA	NA	+	NA	NA
Social studies	NA	NA	NA	NA	+/NR	NA
Algebra I	NA	NA	NA	NA	NA	+
Biology	NA	NA	NA	NA	NA	+
English I	NA	NA	NA	NA	NA	+
English II	NA	NA	NA	NA	NA	+
U.S. History	NA	NA	NA	NA	NA	+/NR

Source. PPfT appraisal results from 2017–2018 and SAS EVAAS teacher-level value-added scores for 2017–2018.
Note. + indicates significant positive associations of PPfT final ratings with student growth measures. NR indicates no relationship between PPfT final ratings and student growth measures. +/NR indicates mixed results across growth measures and correlation statistics. NA indicates grade and subject combination is not applicable due to either no testing in that grade for that subject or no prior testing history in the subject from which to compute the student growth measure correlate (i.e., writing is tested in grade 4, but there is not enough of a STAAR testing history prior to 4th grade from which to derive growth in writing into grade 4).

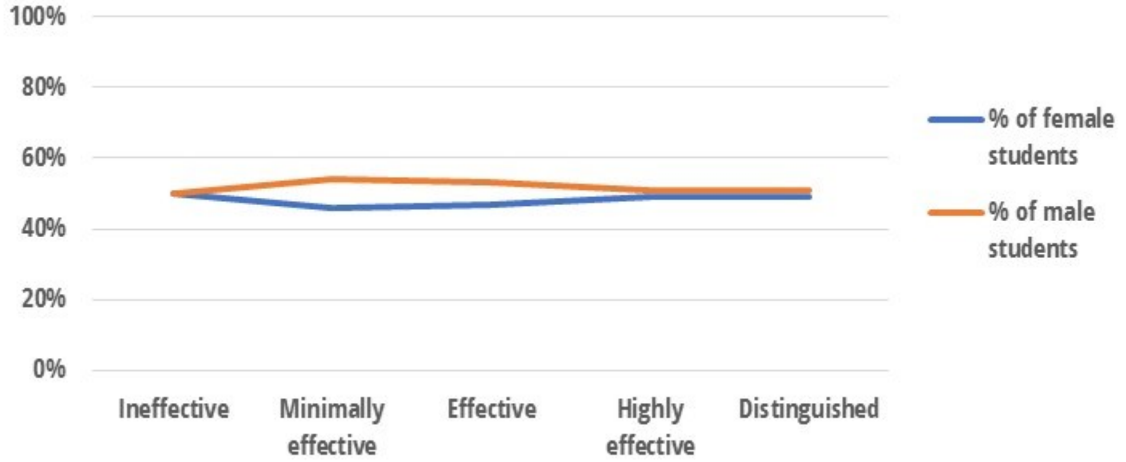
Discriminant Validity: To what extent were teachers' final ratings on PPfT associated with their students' characteristics?

The assessment of discriminant validity is conceptually opposite that of convergent validity. In this study, it was assumed that quality teaching should not be associated with characteristics of the students served by the teachers. Therefore, teachers' PPfT final ratings should not correlate with their students' characteristics. Discriminant validity was examined with the subsets of students served by each teacher. The student characteristics examined included gender, limited English proficiency (LEP) status, economic disadvantage (ECONDIS) status, gifted and talented (GT) status, special education (SPED) status, and race/ethnicity.

The gender of the students served by teachers appeared to operate independently of the final ratings teachers received (Figure 4, Appendix B). Trend-wise, the percentage of male students served by teachers decreased slightly for higher PPfT final ratings, compared with lower PPfT final ratings, and vice versa for the percentage of female students served by teachers. However, actual gender differences were negligible for most PPfT final ratings, and the overall change in percentages of students across rating categories was nearly trivial for gender.

The gender of the students served by teachers appeared to operate independently of the final ratings teachers received.

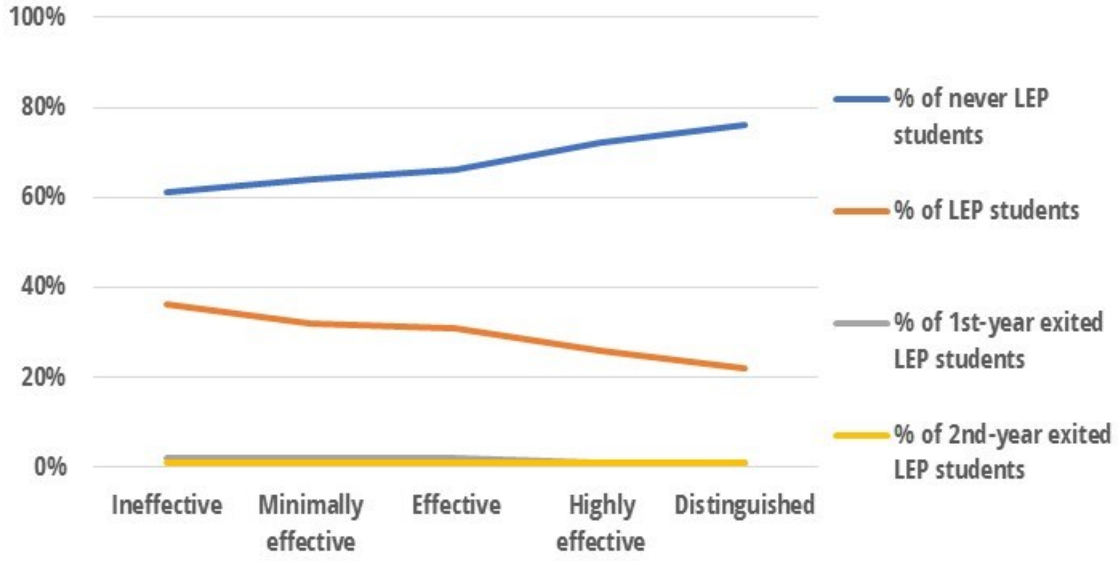
Figure 4.
Student gender was not associated with teachers' PPfT final ratings.



Source. PPfT appraisal results from 2017–2018 and student demographic information from 2017–2018 Texas Student Data System (TSDS).
Note. The correlation coefficient for the percentage of female students and the PPfT final rating was *rho* = 0.04 (p = 0.007). The correlation coefficient for the percentage of male students and the PPfT final rating was *rho* = -0.04 (p = 0.007).

The LEP status of the students served by teachers appeared to operate in some dependency with the final ratings teachers received (Figure 5, Appendix B). The smaller the percentage of LEP students served by teachers, the higher teachers' PPfT final ratings. The greater the percentage of never-LEP students served by teachers, the greater teachers' PPfT final ratings. No relationships were observed for the percentages of 1st- and 2nd-year exited LEP students served and teachers' PPfT final ratings.

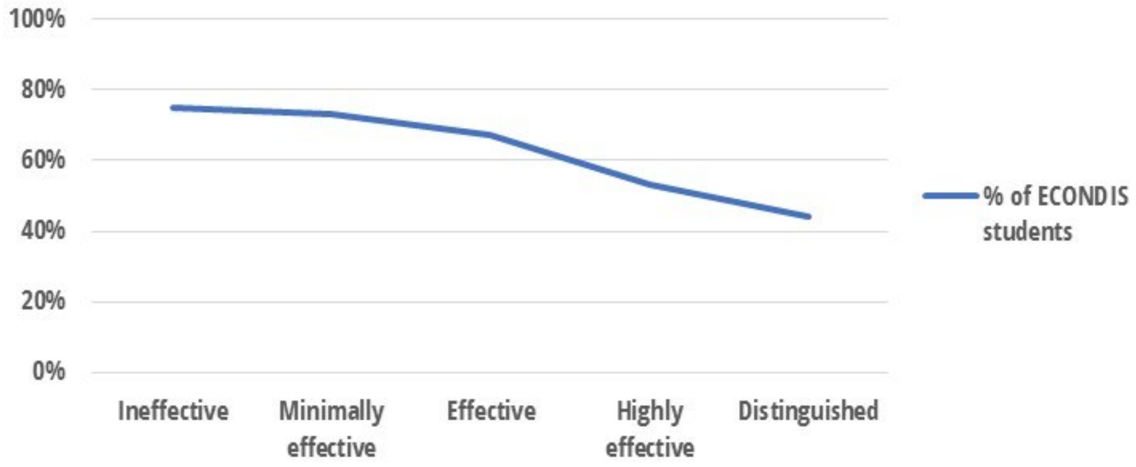
Figure 5.
Student LEP status was associated with teachers' PPfT final ratings.



Source. PPfT appraisal results from 2017–2018 and student demographic information from 2017–2018 Texas Student Data System (TSDS).
Note. The correlation coefficient for the percentage of never-LEP students and the PPfT final rating was *rho* = 0.12 (p < 0.001). The correlation coefficient for the percentage of LEP students and PPfT final rating was *rho* = -0.11 (p < 0.001). The correlation coefficient for the percentage of 1st-year exited LEP students and the PPfT final rating was *rho* = -0.11 (p < 0.001). The correlation coefficient for the percentage of 2nd-year exited LEP students and the PPfT final rating was *rho* = -0.06 (p < 0.001).

The ECONDIS status of the students served by teachers appeared to operate in some dependency with the PPfT final ratings teachers received (Figure 6, Appendix B). The smaller the percentage of ECONDIS students served by teachers, the higher teachers' PPfT final ratings.

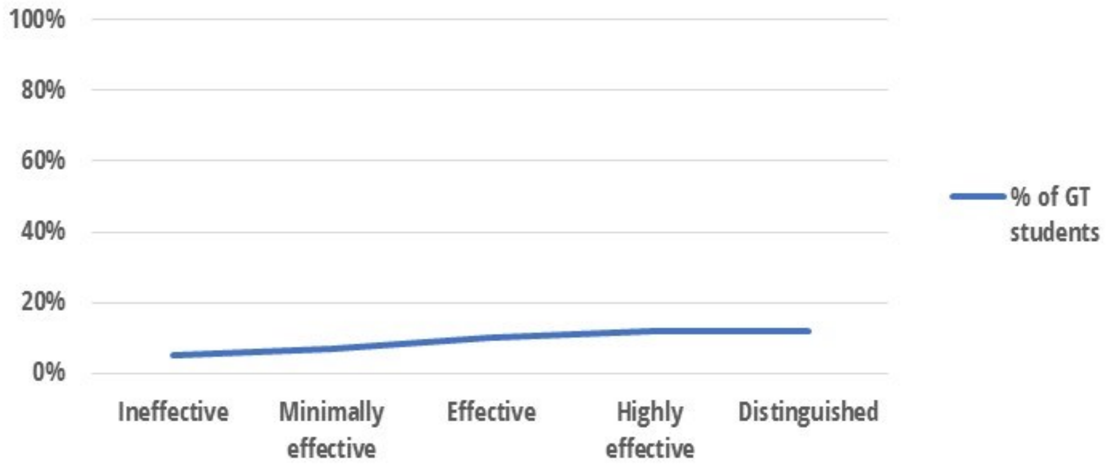
Figure 6.
Student ECONDIS status was associated with teachers' PPfT final ratings.



Source. PPfT appraisal results from 2017–2018 and student demographic information from 2017–2018 Texas Student Data System (TSDS).
Note. The correlation coefficient for the percentage of ECONDIS students and the PPfT final rating was *rho* = -0.24 (p < 0.001).

The GT status of the students served by teachers appeared to operate independently of the final ratings teachers received (Figure 7, Appendix B). Although a greater percentage of GT students served by teachers was associated with a slight increase in teachers' PPfT final ratings, the overall increase in percentages of students by GT status across PPfT final ratings was not large enough to conclude a meaningful relationship existed between the student characteristic and PPfT final rating.

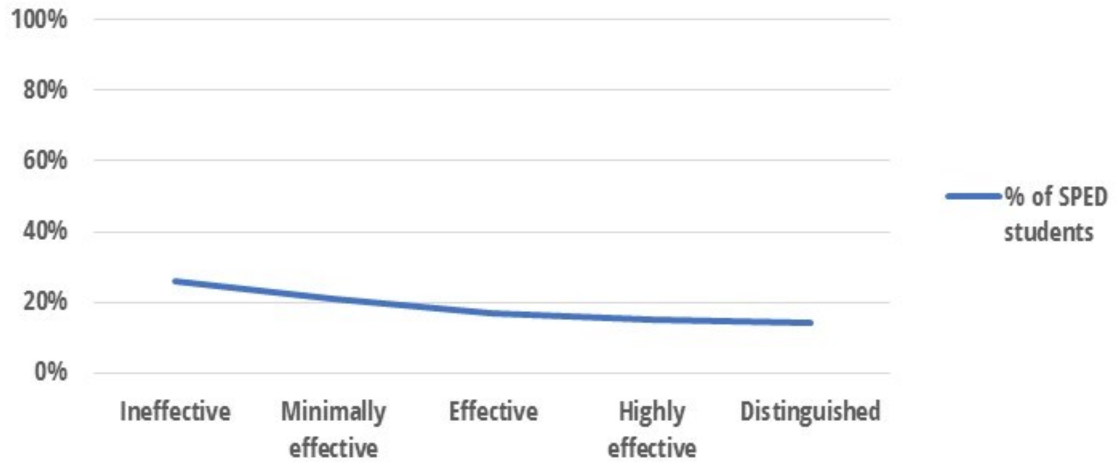
Figure 7.
Student GT status was not associated with teachers' PPfT final ratings.



Source. PPfT appraisal results from 2017–2018 and student demographic information from 2017–2018 Texas Student Data System (TSDS).
Note. The correlation coefficient for the percentage of GT students and the PPfT final rating was *rho* = 0.10 (p < 0.001).

Similar to GT status, the SPED status of the students served by teachers appeared to operate independently of the final ratings teachers received (Figure 8, Appendix B). Although a smaller percentage of SPED students served by teachers was associated with a slight increase in teachers' PPfT final ratings, the overall decrease in percentages of students by SPED status across PPfT final ratings was not large enough to conclude a meaningful relationship existed between the student characteristic and PPfT final rating.

Figure 8.
Student SPED status was not associated with teachers' PPfT final ratings.

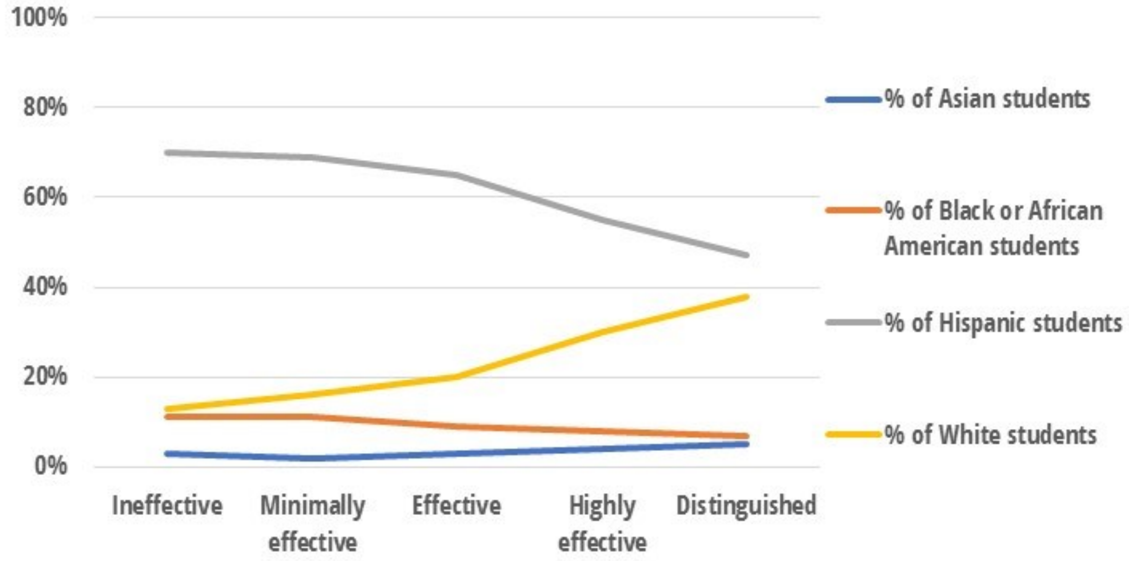


Source. PPfT appraisal results from 2017–2018 and student demographic information from 2017–2018 Texas Student Data System (TSDS).
Note. The correlation coefficient for the percentage of SPED students and the PPfT final rating was *rho* = -0.09 (p < 0.001).

The relationship between teachers' PPfT final ratings and their students' race/ethnicity differed by race/ethnicity category (Figure 9, Appendix B). No meaningful trends were observed for the percentages of Asian and Black or African American students associated with teachers' PPfT final ratings. However, associations were observed with the percentages of Hispanic and White students. The greater the percentage of Hispanic students served by teachers, the lower teachers' PPfT final ratings. The greater the percentage of White students served by teachers, the higher teachers' PPfT final ratings. In other words, teachers with high PPfT final ratings tended to have fewer Hispanic students than did teachers with low PPfT final ratings, and teachers with high PPfT final ratings tended to have more White students than did teachers with low PPfT final ratings.

Teachers' PPfT final ratings were associated with the percentages of Hispanic and White students served by the teachers, but not the percentages of African American or Asian students.

Figure 9.
Student The percentages of White and Hispanic students served were associated with teachers' PPfT final ratings, but the percentages of Asian and African American were not.



Source. PPfT appraisal results from 2017–2018 and student demographic information from 2017–2018 Texas Student Data System (TSDS).
Note. The correlation coefficient for the percentage of Asian students and the PPfT final rating was *rho* = 0.1 (*p* < 0.001). The correlation coefficient for the percentage of Black of African American students and the PPfT final rating was *rho* = -0.11 (*p* < 0.001). The correlation coefficient for the percentage of Hispanic students and the PPfT final rating was *rho* = -0.24 (*p* < 0.001). The correlation coefficient for the percentage of White students and the PPfT final rating was *rho* = 0.25 (*p* < 0.001).

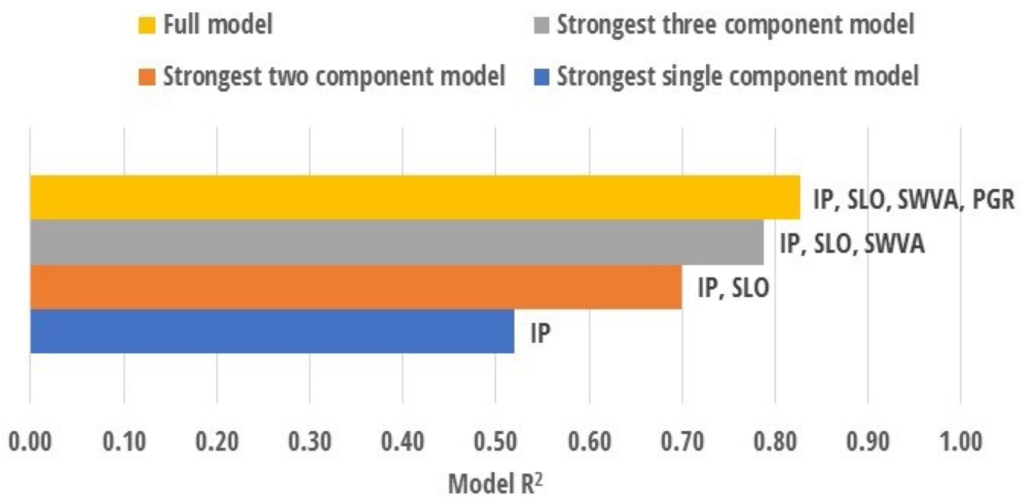
Dominance: What components of PPfT were most important to the prediction of final ratings?

Dominance analysis is a way of assessing the relative importance of the components of the PPfT appraisal in producing the final rating (Azen & Budescu, 2003). PPfT final ratings were based on an overall summative score, and the overall summative score was based on points accumulated from the four components of the PPfT appraisal. Each of the PPfT appraisal components (IP, PGR, SLOs, and SWVA) was weighted in its contribution to the PPfT summative score (50%, 25%, 15%, and 10%, respectively, for the standard PPfT appraisal plan). However, the components' relative importance in a multiple regression model predicting final ratings can differ from the importance assigned by weights, depending on the variance of each measure, its strength of association with the PPfT final rating, and the strength of association between the components of the appraisal.

Results of the dominance analysis revealed that IP ratings were the most important contributor to predicting final ratings (conditional and general dominance), then SLO ratings, followed by PGR ratings, and lastly SWVA ratings (Appendix C). The reversal of SLO and PGR dominance in predicting PPfT final ratings, relative to their importance assigned by weights for scoring in appraisal computations, was particularly interesting. The switch in importance when shifting from scoring weights to predictive contribution emphasized the importance of the distribution quality for appraisal components. Although weighted differently, the similar, skewed distributions of IP and PGR shown in Figure 2, coupled with the dominance analysis, suggests that very little additional information is being added by PGR and IP ratings over IP ratings alone.

Although the dominance analysis focused on the additional contribution each component made to predicting PPfT final ratings, the analysis also revealed which components were the best predictors in one, two, and three component combinations. For single component models, the R2 for IP was the greatest between all components (Figure 10). When examining pairs of components, IP and SLOs together were the strongest pair of predictors of PPfT final ratings. IP, SLOs, and SWVA had the highest R2 between the three component models.

Figure 10.
PGR provided little unique information in the prediction of PPfT final ratings.



Source. PPfT appraisal results from 2017–2018.
Note. See Appendix C for complete dominance analysis.

Interrater Reliability: What was the interrater reliability for teacher instructional practice ratings?

Of the four components of the PPfT appraisal, only IP is scored by two raters. Therefore, the question of interrater reliability could only be asked about instructional practice ratings. Ideally, different raters would provide ratings of the same teacher at the same point in time (i.e., two raters observe a teacher simultaneously), so any differences in ratings received by the teacher could be attributed to differences in how the raters scored the teacher. Under a perfect system, raters consistently observe and score the same thing in the same way at the same time, and teachers use fall feedback to improve their craft throughout the year and receive higher scores from two raters in the spring observation. In PPfT, two different raters observe every appraised teacher, but they do so at different points in time (i.e., one rater observes in the fall, and a different rater observes in the spring). Consequently, any differences in ratings received by the teacher present a confound between raters and time.

Assessment of interrater reliability suggested adequate agreement between raters but ultimately yielded inconclusive results due to the confounds between raters and time and teacher improvement. Teachers consistently received the same ratings or slightly higher ratings in the spring from rater 2 than they did in the fall from rater 1. Because of this, interrater reliability analyses detected the small deviations between raters and returned results suggesting moderate agreement. Analysis of the difference between fall/rater 1 and spring/rater 2 revealed that ratings on all strands significantly improved from fall to spring (on average by 0.11 of a rating point), and correlation analyses revealed significant positive associations between strand pairs from fall to spring. Taken together, the analyses show a consistent pattern of improvement for teachers between fall/rater 1 and spring/rater 2. Although consistent, the improvement meant absolute agreement between raters could not be expected. Interrater reliability analysis using Cohen's weighted kappa coefficient revealed a moderate degree of agreement between raters on all strands (Table 3).

Table 3.
Fall to Spring Comparisons showed fair to moderate rater agreement using Cohen's weighted kappa.

IP strand	Overall (*N* = 5,577)	ES (*n* = 3,004)	MS (*n* = 1,082)	HS (*n* = 1,339)	Special (*n* = 152)
Student engagement	*wk* = .44	*wk* = .43	*wk* = .49	*wk* = .41	*wk* = .42
Assessment and feedback	*wk* = .4	*wk* = .41	*wk* = .41	*wk* = .37	*
Differentiation	*wk* = .37	*wk* = .38	*wk* = .37	*wk* = .33	*
PS and CT	*wk* = .4	*wk* = .4	*wk* = .44	*wk* = .39	*
Classroom expectations	*wk* = .42	*wk* = .42	*wk* = .39	*wk* = .4	*
Routines and procedures	*wk* = .45	*wk* = .49	*wk* = .45	*wk* = .35	*
Classroom climate	*wk* = .45	*wk* = .45	*wk* = .47	*	*

Source. PPfT appraisal results from 2017–2018.
Note. *wk* indicates Cohen's weighted kappa coefficient. All comparisons were significant, rejecting the null hypothesis of no agreement. * indicates analysis could not be run because the contingency table was not square (e.g., spring ratings often did not include all rating values, usually 1s, therefore the contingency table was not square and kappa was not run by SAS EG). Paired samples *t* tests revealed that ratings on all strands significantly improved from fall to spring, on average by 0.11 of a rating point. Pearson correlation coefficients varied from 0.45 for differentiation to 0.53 for routines and procedures. Spearman correlation coefficients varied from 0.44 to 0.50 for differentiation and routines and procedures, respectively.

Internal Consistency: To what extent were strand ratings within components correlated, and to what extent were the components ratings of PPfT correlated?

If we assume that the four components of the PPfT final rating collectively measured the same construct (i.e., quality teaching), and similarly that the strands of IP and PGR each measured the same thing, then we would expect consistent patterns between the ratings teachers received on the components to judge each as internally consistent. The internal consistency of ratings was assessed separately on four sets of measures: (a) the seven strands of IP rated in the fall, (b) the seven strands of IP rated in the spring, (c) the five strands of PGR, and (d) overall scores on the four components of PPfT final ratings.

Fall IP ratings, spring IP ratings, and PGR ratings all showed similar evidence of strong internal consistency. Items within each set showed significant positive correlations. All sets showed large (e.g., ~0.8) standardized Cronbach's alpha coefficients. On all three correlation sets, Cronbach's alpha could not be improved with the removal of any item from the set of items (Table 4, Table 5, and Table 6).

Table 4.
Fall IP strands showed strong internal consistency.

Overall standardized Cronbach's alpha coefficient	Deleted strand	Adjusted standardized Cronbach's alpha coefficient with deletion
0.873	Fall student engagement	0.846
	Fall assessment and feedback	0.856
	Fall differentiation	0.861
	Fall PS and CT	0.858
	Fall classroom expectations	0.850
	Fall routines and procedures	0.855
	Fall classroom climate	0.855

Source. PPfT appraisal results from 2017–2018.
Note. The seven strands of instructional practice from the fall observation were all significantly positively correlated between all 5,577 teachers appraised under PPfT in 2017–2018. Correlation coefficients varied from the smallest at 0.41 to the largest at 0.59. The standardized Cronbach's alpha coefficient was 0.873 for the set of ratings. The set of fall measures was further explored to see if the standardized Cronbach's alpha coefficient would improve with the removal of any single strand from the set of seven IP strands. In all cases of strand removal, alpha decreased, suggesting that the original set of seven fall strands are a more internally consistent set of measures than any subset of fewer than the seven fall strands.

Table 5.
Spring IP strands showed strong internal consistency.

Overall standardized Cronbach's alpha coefficient	Deleted strand	Adjusted standardized Cronbach's alpha coefficient with deletion
0.865	Spring student engagement	0.836
	Spring assessment and feedback	0.847
	Spring differentiation	0.855
	Spring PS and CT	0.852
	Spring classroom expectations	0.841
	Spring routines and procedures	0.845
	Spring classroom climate	0.845

Source. PPfT appraisal results from 2017–2018.
Note. The seven strands of instructional practice for the spring observation were all significantly positively correlated between all 5,577 teachers appraised under PPfT in 2017–2018. Correlation coefficients varied from the smallest at 0.40 to the largest at 0.57. The standardized Cronbach's alpha coefficient was 0.865 for the set of ratings. The set of spring measures was further explored to see if the standardized Cronbach's alpha coefficient would improve with the removal of any single strand from the set of seven strands. In all cases of strand removal, alpha decreased, suggesting that the original set of seven spring IP strands is a more internally consistent set of measures than is any subset of fewer than seven spring strands.

Table 6.
PGR strands showed strong internal consistency.

Overall standardized Cronbach's alpha coefficient	Deleted strand	Adjusted standardized Cronbach's alpha coefficient with deletion
0.810	Professional development and reflection	0.771
	Collaboration and contributions	0.768
	Lesson planning and data use	0.763
	Relational communication	0.780
	compliance	0.785

Source. PPfT appraisal results from 2017–2018.
Note. The five strands of PGR were all significantly positively correlated between all 5,577 teachers appraised under PPfT in 2017–2018. Correlation coefficients varied from the smallest at 0.40 to the largest at 0.54. The standardized Cronbach's alpha coefficient was 0.810 for the set of ratings. The set of PGR measures was further explored to see if the standardized Cronbach's alpha coefficient would improve with the removal of any single strand from the set of five strands. In all cases of strand removal, alpha decreased, suggesting that the original set of five PGR strands is a more internally consistent set of measures than any subset of fewer than five PGR strands.

The appraisal components of PPfT final ratings showed evidence of somewhat weak internal consistency (Table 7). The four components of PPfT final ratings were all significantly positively correlated, but the standardized Cronbach's alpha coefficient was relatively small (i.e., alpha = 0.511) for the set of ratings. An alpha of that size, although still acceptable, indicates poor internal consistency. In one case, component removal resulted in an increase in alpha, with the exclusion of the SWVA component; however, internal consistency would only improve to a standardized Cronbach's alpha coefficient of 0.574 up from 0.511, a very small change that would not affect the overall conclusion of poor internal consistency.

Table 7.
PPfT appraisal components showed acceptable, yet weak internal consistency.

Overall standardized Cronbach's alpha coefficient	Deleted strand	Adjusted standardized Cronbach's alpha coefficient with deletion
0.511	IP rating	0.341
	PGR rating	0.302
	SLO rating	0.495
	SWVA Rating	0.574

Source. PPfT appraisal results from 2017–2018.
Note. The four components of PPfT final ratings were all significantly positively correlated for the 4,515 teachers appraised on a standard plan under PPfT in 2017–2018. Correlation coefficients varied from the smallest at 0.04 (SWVA and IP) to the largest at 0.58 (PGR and IP). The standardized Cronbach's alpha coefficient was 0.511 for the set of ratings. Removal of the SWVA component improved standardized Cronbach's alpha coefficient (i.e., 0.574 up from 0.511).

So what can be concluded about whether the components reliably measure the same construct (i.e., quality teaching)? It is worth unpacking the findings further, given the off nominal results of the internal consistency analysis of the PPfT appraisal components (i.e., internal consistency between appraisal components was not as strong as it was between the strand ratings of IP and PGR). There are several considerations to weigh.

First, the correlation matrix of the PPfT appraisal components facilitates seeing what is going on (Table 8). When focusing on the strength of association between components, essentially IP and PGR are the only components that share a strong relationship. In other words, teachers' IP ratings are positively associated with their PGR ratings, but not with other components. Teachers' SLO ratings do not share a strong association with any components, nor do teachers' SWVA ratings.

Table 8.
Between the components of PPfT final ratings, only IP and PGR share a strong relationship.

	Pearson correlation coefficients and p values (*N* = 4,515)			
	IP rating	PGR rating	SLO rating	SWVA rating
IP rating	1	0.58 *p* < 0.001	0.18 *p* < 0.001	0.04 *p* = 0.01
PGR rating	0.580 *p* < 0.001	1	0.17 *p* < 0.001	0.12 *p* < 0.001
SLO rating	0.18 *p* < 0.001	0.17 *p* < 0.001	1	0.16 *p* < 0.001
SWVA rating	0.04 *p* = 0.01	0.12 *p* < 0.001	0.16 *p* < 0.001	1

Source. PPfT appraisal results from 2017–2018.

Second, IP and PGR share similar distinctions from SLOs and SWVA both conceptually and procedurally. Conceptually, IP and PGR are measures of the teacher, while SLOs and SWVA are measures of student growth. Process-wise, IP and PGR data are dependent upon school leadership (i.e., ratings), but SLO and SWVA data operate independently of school leadership. SLO data are provided by the teacher through the SLO process, and SWVA data are modeled based on the average growth of students at a school, compared with the average growth of peers in the state. Consequently, appraisal components may measure quality teaching, but two different aspects of quality teaching. Exploratory factor analysis on the four components of PPfT final ratings suggests a two-factor solution, one with IP and PGR and another with the two student growth components (Appendix D). The two-factor solution may represent a system that collectively measures teaching quality, a construct captured with a measure of teacher efficacy (IP and PGR) and a measure of student growth (SLOs and SWVA).

Lastly, it is worth considering the difference between collective and individual measurement of student growth. SWVA is a measure of collective attribution to student growth, as opposed to the SLO measure of individual attribution to student growth. SWVA is an appraisal component that reflects district values around *our* students versus *my* students. It may not be reasonable to expect that SWVA varies with individual teacher ratings, and this is supported in the internal consistency analysis (i.e., low overall internal consistency and a small improvement of internal consistency with the removal of SWVA). It may be more reasonable to expect that SLOs vary with individual teacher ratings. However, there is poor internal consistency for a three-component model with only IP, PGR, and SLO, and the standardized Cronbach's alpha coefficient for a three-component model with only IP, PGR, and SLOs is improved to 0.734 with the removal of the SLO component.

When SWVA is replaced with mean teacher value-added rating (an individual measure of student growth), three effects are observed in the analyses. First, the correlations between components improve. The correlation of mean teacher value-added ratings with IP and PGR ratings are comparable to or superior to those of SLO ratings with IP and PGR ratings. Second, internal consistency improves (standardized Cronbach's alpha coefficient = 0.62) and alpha does not improve with removal of any components. Lastly, rerunning exploratory factor analysis with the mean teacher value-added ratings instead of SWVA ratings results in a single factor solution (Appendix D).

Summary and Recommendations

Validity and Reliability Summaries

Content Validity Summary

Evidence suggests strong content validity around the entire instructional practices observation, rating, and formative feedback process. However, stakeholders seemed divided on whether they disagree, agree, or don't know if the appraisal system measures teaching quality. Issues around item design (i.e., first person versus third person) and education or uncertainty were considered.

Concurrent Validity Summary

The five rating categories used in PPfT final ratings (i.e., distinguished, highly effective, effective, minimally effective, and ineffective) appear to provide an overall differentiation of teaching quality. However, the highly effective final rating category may not provide an efficient differentiation between the large number of teachers rated highly effective (i.e., 57%, n = 3,155). Between the individual PPfT appraisal components, differentiation of teachers was strongest with SWVA and SLOs and weakest with IP and PGR.

Convergent Validity Summary

In general, correlation analyses showed that as teaching quality increased, so did student growth. In other words, high-quality teaching was associated with greater student growth than was low-quality teaching. Comparisons within test, grade, and subject demonstrated evidence for convergent validity in grades 4 through 7 in all STAAR tested subjects and on the algebra I, biology, English I, and English II, STAAR EOCs. Convergent validity was generally not observed in grade 8 (i.e., STAAR math, reading, social studies, science) and on the U.S. history STAAR EOC.

Discriminant Validity Summary

Overall, results of correlation analyses revealed mixed findings across the various student characteristics observed. The gender of the students served by teachers, GT status, and SPED status appeared to operate independently of the final ratings teachers received. However, the LEP status, ECONDIS status, and the race/ethnicity of the students served by teachers appeared to operate in some dependency with the final ratings teachers received. Teachers with high PPfT final ratings tended to have fewer LEP students, fewer ECONDIS students, fewer Hispanic students, and more White students than did teachers with low PPfT final ratings.

Dominance Summary

Results of dominance analysis revealed that IP ratings were the most important contributor to predicting final ratings, followed by SLO ratings, PGR ratings, and lastly SWVA ratings. Results further underscored the importance of distribution quality (e.g., variance and normality) for each PPfT appraisal component. The distribution of PGR ratings (weighted at 25% of PPfT final rating) contributed less to prediction of the PPfT final rating than did SLO ratings (weighted at 15% of PPfT final rating), and in some cases, less than SWVA ratings (weighted at only 10% of PPfT final rating). Analyses suggest that very little additional information is being added by PGR and IP ratings over IP ratings alone.

Interrater Reliability Summary

Because PPfT IP observations relied on a rating from one rater in the fall and a second rating from a different rater in the spring, interrater reliability analyses had to contend with a confound between raters and time (in addition to explicit teacher use of feedback from fall/rater 1 to improve for the spring/rater 2 observation). Difference, association, and agreement between fall/rater 1 and spring/rater 2 were used together to assess interrater reliability. On average, all teachers improved slightly from their first to their second observation, but ratings remained in moderate agreement. Collectively, the analyses of interrater reliability were interpreted to show evidence of adequate agreement between raters, but ultimately assessment of interrater reliability was inconclusive due to the confounds between raters and time and teacher improvement.

Internal Consistency Summary

Fall IP ratings, spring IP ratings, and PGR ratings all showed evidence of strong internal consistency with significant positive correlations between rated strands and large standardized Cronbach's alpha coefficients. On all three correlation sets, Cronbach's alpha could not be improved with the removal of any item from the set of items. Appraisal components showed evidence of somewhat weak internal consistency. Although the four components of PPfT final ratings were also significantly positively correlated, the standardized Cronbach's alpha coefficient was relatively small, indicating somewhat poor internal consistency. Although internal consistency did not meaningfully improve with removal of any components, exploratory analysis considering replacement of the SWVA component with a teacher value-added component did meaningfully improve internal consistency.

Recommendations

Increase education efforts around measurement of quality teaching.

Although analysis of content validity revealed mixed perceptions about whether the appraisal system measures teaching quality, convergent validity analyses reveal strong associations between teacher PPfT final ratings and the growth of the students served by the teachers. Improvements in survey item design and additional education efforts inclusive of the convergent validity results may help position stakeholders' perceptions more conclusively.

Work on calibrating raters to what 2s and 3s looks like in the classroom.

Although raters seem to be operating with adequate calibration, training has historically emphasized differentiating 3s from 4s. This differentiation between 3s and 4s is successfully evident in the assessment of concurrent validity, but the results also suggest a potential anchoring effect of raters on 3s as their floor rating, given the training emphasis on 3s and 4s. Future training and calibration offerings should attempt to reorient raters to what 2s and 3s look like in the classroom in an effort to improve the quality of the distribution of ratings, thus also improving the potential contribution of PGR ratings to prediction of PPfT final rating.

Leverage the strategic recruiting and compensation to place highly-quality teachers in populations of underserved students.

Leverage the strategic recruiting and compensation of the Comprehensive Schools Improvement Model to address discriminant validity. Use of the strategic recruiting and compensation may help place more highly effective and distinguished teachers with populations of underserved students (i.e., LEP students, ECONDIS students, and Hispanic students).

Partner peer observers with school administrators for instructional practice observations.

Use of floating trained peer observers to partner with school administrators during both observations would eliminate the confound between raters and time. Peer observers have less implicit bias toward the teachers than do the teachers' direct supervisors (i.e., their school administrators). Use of an impartial partner observer may help prevent school administrators from unconsciously rating their own teachers higher than they would an unknown teacher, improve the evaluation of interrater reliability at a campus level, and help connect teacher populations across the district for analysis of systematic bias by content area.

Revisit district values on individual versus collective measurement of student growth in teacher appraisal.

Two possible perspectives on the components of teacher appraisal were explored through the internal consistency analyses. One perspective acknowledges the district values around the collective contribution to the growth of all students at a teacher's school, accepts the resulting lower internal consistency of appraisal components with inclusion of the collective measure, and presents a possible two-factor system (teacher quality and student growth) for measuring teaching quality comprising the current four appraisal components. The other perspective focuses instead on improving internal consistency through replacement of the collective SWVA measure with an individual mean teacher value-added measure and presents a single-factor solution to measuring teaching quality comprising a teacher value-added measure of student growth, IP, PGR, and SLOs.

Appendix A

To examine convergent validity, the correlations between PPfT final ratings (using a 5-point ordinal scale where 1 = ineffective and 5 = distinguished) and student growth measures were examined. A set of student growth metrics based on STAAR 3 through 8 and EOC assessments were used as correlates with PPfT final ratings and examined within test, grade, and subject. The set of student growth metrics included the EVAAS growth measure, growth index, and growth level, all at the individual-teacher level. Smaller subsets of appraised teachers were examined in this analysis because not all teachers had individual growth measures. Student growth was calculated for the students served by the teacher and weighted by the percentage of time the teacher shared the student. Results of both Pearson and Spearman correlations are shown by grade and subject in Tables A.1 through A.4 in Appendix A.

In both math and reading, significant positive correlations between PPfT final ratings and the student growth measures were observed in grades 4 through 7, ranging from a low of *r* = 0.21 to a high of *r* = 0.39. The significant positive correlations were not observed in grade 8 for either subject. In science, significant positive correlations between PPfT final ratings and the student growth measures were observed in grade 5, ranging from a low of *r* = 0.27 to a high of *r* = 0.33. The significant positive correlations were not observed in grade 8 for science. In writing, significant positive correlations between PPfT final ratings and the student growth measures were observed in grade 7, ranging from a low of *r* = 0.4 to a high of *r* = 0.45. Findings were mixed in social studies for grade 8. On the algebra I, biology, English I, and English II, EOC significant positive correlations between PPfT final ratings and the student growth measures were observed, ranging from a low of *r* = 0.25 to a high of *r* = 0.52. Findings were mixed on the U.S. history EOC.

Table A.1
Math Correlations

Grade	Pearson correlation coefficients and p values			Spearman correlation coefficients and p values		
	Teacher gain	Teacher comparison index	Teacher growth level	Teacher gain	Teacher comparison index	Teacher growth level
4 (*n* = 230)	0.30 *p* < 0.001	0.29 *p* < 0.001	0.28 *p* < 0.001	0.29 *p* < 0.001	0.28 *p* < 0.001	0.28 *p* < 0.001
5 (*n* = 144)	0.39 *p* < 0.001	0.37 *p* < 0.001	0.37 *p* < 0.001	0.38 *p* < 0.001	0.37 *p* < 0.001	0.37 *p* < 0.001
6 (*n* = 72)	0.220 *p* = 0.06	0.210 *p* = 0.08	0.285 *p* = 0.02	0.265 *p* = 0.02	0.253 *p* = 0.03	0.263 *p* = 0.03
7 (*n* = 60)	0.34 *p* = 0.009	0.33 *p* = 0.01	0.32 *p* = 0.012	0.32 *p* = 0.014	0.33 *p* = 0.009	0.31 *p* = 0.016
8 (*n* = 59)	0.22 *p* = 0.089	0.22 *p* = 0.092	0.22 *p* = 0.096	0.27 *p* = 0.037	0.18 *p* = 0.181	0.19 *p* = 0.154

Source. PPfT appraisal results from 2017–2018 and SAS EVAAS teacher-level value-added scores for 2017–2018.

Table A.2

Reading Correlations

Grade	Pearson correlation coefficients and p values			Spearman correlation coefficients and p values		
	Teacher gain	Teacher comparison index	Teacher growth level	Teacher gain	Teacher comparison index	Teacher growth level
4 (*n* = 259)	0.35 *p* < 0.001	0.34 *p* < 0.001	0.29 *p* < 0.001	0.33 *p* < 0.001	0.33 *p* < 0.001	0.27 *p* < 0.001
5 (*n* = 165)	0.24 *p* = 0.002	0.25 *p* = 0.002	0.18 *p* = 0.021	0.25 *p* < 0.001	0.24 *p* = 0.002	0.18 *p* = 0.021
6 (*n* = 91)	0.46 *p* < 0.001	0.41 *p* < 0.001	0.37 *p* < 0.001	0.42 *p* < 0.001	0.41 *p* < 0.001	0.35 *p* < 0.001
7 (*n* = 73)	0.26 *p* = 0.024	0.25 *p* = 0.033	0.22 *p* = 0.065	0.23 *p* = 0.049	0.24 *p* = 0.037	0.18 *p* = 0.129
8 (*n* = 72)	0.14 *p* = 0.236	0.08 *p* = 0.501	0.12 *p* = 0.323	0.08 *p* = 0.524	0.06 *p* = 0.603	0.07 *p* = 0.571

Source. PPfT appraisal results from 2017–2018 and SAS EVAAS teacher-level value-added scores for 2017–2018.

Table A.3

Science, Writing, and Social Studies Correlations

Subject	Grade	Pearson correlation coefficients and p values			Spearman correlation coefficients and p values		
		Teacher gain	Teacher comparison index	Teacher growth level	Teacher gain	Teacher comparison index	Teacher growth level
Science	5 (*n* = 147)	0.33 *p* < 0.001	0.32 *p* < 0.001	0.29 *p* < 0.001	0.30 *p* < 0.001	0.30 *p* < 0.001	0.27 *p* = 0.001
	8 (*n* = 50)	0.15 *p* = 0.286	0.15 *p* = 0.285	0.15 *p* = 0.295	0.18 *p* = 0.212	0.19 *p* = 0.186	0.16 *p* = 0.261
Writing	7 (*n* = 64)	0.42 *p* = 0.001	0.44 *p* < 0.001	0.46 *p* < 0.001	0.40 *p* = 0.001	0.43 *p* < 0.001	0.45 *p* < 0.001
Social studies	8 (*n* = 53)	0.28 *p* = 0.039	0.32 *p* = 0.020	0.27 *p* = 0.050	0.26 *p* = 0.056	0.27 *p* = 0.049	0.25 *p* = 0.075

Source. PPfT appraisal results from 2017–2018 and SAS EVAAS teacher-level value-added scores for 2017–2018.

Table A.4

EOC Correlations

Grade	Pearson correlation coefficients and p values			Spearman correlation coefficients and p values		
	Teacher gain	Teacher comparison index	Teacher growth level	Teacher gain	Teacher comparison index	Teacher growth level
Algebra I (*n* = 102)	0.35 *p* < 0.001	0.36 *p* < 0.001	0.25 *p* = 0.011	0.25 *p* = 0.011	0.25 *p* = 0.011	0.23 *p* = 0.022
Biology (*n* = 58)	0.42 *p* = 0.001	0.40 *p* = 0.002	0.30 *p* = 0.024	0.43 *p* = 0.001	0.40 *p* = 0.002	0.33 *p* = 0.013
English I (*n* = 70)	0.49 *p* < 0.001	0.49 *p* < 0.001	0.35 *p* = 0.003	0.52 *p* < 0.001	0.52 *p* < 0.001	0.35 *p* = 0.003
English II (*n* = 60)	0.35 *p* = 0.007	0.27 *p* = 0.034	0.28 *p* = 0.033	0.29 *p* = 0.023	0.25 *p* = 0.057	0.26 *p* = 0.047
U.S. history (*n* = 51)	0.32 *p* = 0.024	0.30 *p* = 0.031	0.22 *p* = 0.117	0.29 *p* = 0.039	0.27 *p* = 0.057	0.22 *p* = 0.128

Source. PPfT appraisal results from 2017–2018 and SAS EVAAS teacher-level value-added scores for 2017–2018.

Appendix B

Discriminant validity was examined for the subset of teachers with a final 2017–2018 PPfT appraisal and course information in the summer 2018 PEIMS Staff Teaching Class Assignments file for the 2017–2018 school year. Students were linked to the courses offered by the teachers, and the percentage of each demographic characteristic of the students served by the teachers was calculated. Teachers with less than 10 students were excluded from the analyses. For all discriminant validity analyses, 4,306 teachers were identified in both data systems with 10 or more students. Student characteristics examined included gender, LEP status, ECONDIS status, GT status, SPED status, and race/ethnicity. Two analyses were performed on final rating and student characteristic data.

Correlations between PPfT final ratings (using a 5-point ordinal scale where 1 = ineffective and 5 = distinguished) and the percentages of each characteristic of the students served by each teacher were examined. Analysis of variance was also run comparing the differences in mean percentages between the five final rating categories. Overall, results of correlation analyses revealed mixed findings, dependent upon the specific student characteristics observed. The large sample size (i.e., *N* = 4,306 teachers) resulted in detection (i.e., statistical significance) of some very small, yet not practically meaningful correlations between PPfT final ratings and percentages of student characteristics. Consequently, results of discriminant validity analyses focused on the direction of relationships (i.e., positive or negative trend) and size of relationships (i.e., strength of correlation coefficient and absolute difference between percentages for ineffective and distinguished ratings) in the context of the direction and size of all relationships examined for the student characteristic. Results of each analysis are shown for each student characteristic in Tables B.1 through B.6.

Table B.1
Gender Correlations and Means

Demographic correlate	Correlation results with PPfT final rating rho (p value)	Mean percentages by PPfT final rating (column percent)				
		Ineffective	Minimally effective	Effective	Highly effective	Distinguished
% of female students	0.04 (0.007)	50%	46%	47%	49%	49%
% of male students	-0.04 (0.007)	50%	54%	53%	51%	51%

Source. PPfT appraisal results from 2017–2018 and student demographic information from 2017–2018 Texas Student Data System (TSDS).

Table B.2
LEP Status Correlations and Means

Demographic correlate	Correlation results with PPfT final rating rho (p value)	Mean percentages by PPfT final rating (column percent)				
		Ineffective	Minimally effective	Effective	Highly effective	Distinguished
% of never-LEP students	0.12 (<0.001)	61%	64%	66%	72%	76%
% of LEP students	-0.11 (<0.001)	36%	32%	31%	26%	22%
% of 1st-year exited LEP students	-0.11 (<0.001)	2%	2%	2%	1%	1%
% of 2nd-year exited LEP students	-0.06 (<0.001)	1%	1%	1%	1%	1%

Source. PPfT appraisal results from 2017–2018 and student demographic information from 2017–2018 Texas Student Data System (TSDS).

Table B.3
ECONDIS Status Correlations and Means

Demographic correlate	Correlation results with PPfT final rating rho (p value)	Mean percentages by PPfT final rating (column percent)				
		Ineffective	Minimally effective	Effective	Highly effective	Distinguished
% of ECONDIS students	-0.24 (<0.001)	75%	73%	67%	53%	44%
% of not-ECONDIS students	0.24 (<0.001)	25%	27%	33%	47%	56%

Source. PPfT appraisal results from 2017–2018 and student demographic information from 2017–2018 Texas Student Data System (TSDS).

Table B.4
GT Status Correlations and Means

Demographic correlate	Correlation results with PPfT final rating rho (p value)	Mean percentages by PPfT final rating (column percent)				
		Ineffective	Minimally effective	Effective	Highly effective	Distinguished
% of GT students	0.10 (<0.001)	5%	7%	10%	12%	12%
% of not-GT students	-0.10 (<0.001)	95%	93%	90%	88%	88%

Source. PPfT appraisal results from 2017–2018 and student demographic information from 2017–2018 Texas Student Data System (TSDS).

Table B.5
SPED Status Correlations and Means

Demographic correlate	Correlation results with PPfT final rating rho (p value)	Mean percentages by PPfT final rating (column percent)				
		Ineffective	Minimally effective	Effective	Highly effective	Distinguished
% of SPED students	-0.09 (<0.001)	26%	21%	17%	15%	14%
% of not-SPED students	0.09 (<0.001)	74%	79%	83%	85%	86%

Source. PPfT appraisal results from 2017–2018 and student demographic information from 2017–2018 Texas Student Data System (TSDS).

Table B.5
Race/Ethnicity Correlations and Means

Demographic correlate	Correlation results with PPfT final rating rho (p value)	Mean percentages by PPfT final rating (column percent)				
		Ineffective	Minimally effective	Effective	Highly effective	Distinguished
% of Asian students	0.1 (<.001)	3%	2%	3%	4%	5%
% of Black or African American students	-0.11 (<.001)	11%	11%	9%	8%	7%
% of Hispanic students	-0.24 (<.001)	70%	69%	65%	55%	47%
% of White students	0.25 (<.001)	13%	16%	20%	30%	38%

Source. PPfT appraisal results from 2017–2018 and student demographic information from 2017–2018 Texas Student Data System (TSDS).
Note. Results of the analyses of American Indian or Alaskan Native, Hawaiian or Pacific Islander, and two or more races were excluded due to limited variability. The maximum percentages observed for any teachers were 10%, 6%, and 25%, respectively.

Appendix C

In the PPfT dominance analysis, the contribution of each component to the model's R^2 was assessed for all possible combinations of components: each appraisal component by itself, each possible pair of components, each possible combination of three components, and the full model inclusive of all four components. Three levels of dominance were considered: complete, conditional, and general. The three levels of dominance were related to each other in a hierarchical fashion.

Complete dominance implies conditional dominance, which, in turn, implies general dominance. Complete dominance indicates that the additional contribution of a component to the model's R^2 exceeds that of all other components in all model combinations. Conditional dominance compares the average additional R^2 contribution of each variable within each model size (i.e., k). General dominance compares the overall average R^2 contribution of each variable across all models.

In the k = 0 model, each component was modeled independently as a predictor of PPfT final rating. In the k = 1 model, the additional contribution to R^2 was explored for each component in all possible two component models. In the k = 2 model, the additional contribution to R^2 was explored for each component in all possible three component models. In the k = 3 model, the additional contribution to R^2 was explored for each component in all possible three component models. In the k = 4 model, all four components were modeled together as predictors of PPfT final rating. Examining Table C.1 in the order of predictor importance, we observe that:

For IP ratings (i.e., X_1) versus SLO ratings (i.e., X_3), PGR ratings (i.e., X_2) and SWVA ratings (i.e., X_4):

- IP ratings showed conditional dominance over SLO ratings in all k models and general dominance over SLO ratings; however, IP ratings came close to complete dominance over SLO ratings, with the singular exception of the contribution of IP ratings versus SLO ratings to the PGR ratings only model.
- IP ratings showed complete dominance over PGR ratings.
- IP ratings showed complete dominance over SWVA ratings.

For SLO ratings (i.e., X_3) versus IP ratings (i.e., X_1), PGR ratings (i.e., X_2) and SWVA ratings (i.e., X_4):

- SLO ratings exceeded the individual R^2 contribution of IP ratings to the PGR only model but were otherwise dominated by IP ratings.
- SLO ratings showed conditional dominance over PGR ratings in the k = 2 and k = 3 models, but PGR ratings showed conditional dominance over SLO ratings in the k = 0 and k = 1 models. SLO ratings and PGR ratings were equivalent in their overall average contribution to the full model. SLO ratings also exceeded the individual additional R^2 contribution of PGR ratings to the full model (additional R^2 contribution = 0.1304 versus 0.0389, for SLO ratings and PGR ratings in the k = 3 model, respectively).

- SLO ratings showed complete dominance over SWVA ratings.

For PGR ratings (X_2) versus IP ratings (i.e., X_1), SLO ratings (i.e., X_3), and SWVA ratings (X_4):

- PGR ratings were completely dominated by IP ratings.
- PGR ratings showed conditional dominance over SLO ratings in the k = 0 and k = 1 models, but SLO ratings showed conditional dominance over PGR ratings in the k = 2 and k = 3 models.
- PGR ratings showed general dominance over SWVA ratings.
- PGR ratings showed the smallest individual additional R^2 contribution to the full model out of all four predictors (i.e., 0.0389), even smaller than the contribution of SWVA (i.e., 0.0748).

For SWVA ratings (X_4) versus IP ratings (i.e., X_1), PGR ratings (X_2), and SLO ratings (i.e., X_3):

- SWVA ratings were completely dominated by IP and SLO ratings.
- SWVA ratings were generally dominated by PGR ratings; however, SWVA ratings showed conditional dominance over PGR ratings in the k = 3 model. The additional contribution to R^2 of SWVA ratings to the three-component model (i.e., $X_1X_2X_3$) was greater than the additional contribution to R^2 of PGR ratings to the three-component model (i.e., $X_1X_3X_4$).

The conditional dominance of SLO ratings (weighted at 15% of the appraisal) over PGR ratings (weighted at 25% of the appraisal) in the prediction of the PPfT final rating in the $k = 2$ and $k = 3$ models emphasizes the importance of appraisal component distribution quality. The similar, skewed distributions of IP and PGR shown in Figure 2, coupled with the dominance analysis, suggest that very little additional information was added by PGR and IP ratings over IP ratings alone. Even SWVA ratings (weighted at 10% of appraisal) exceeded the individual additional R^2 contribution to the full model (additional R^2 contribution = 0.0748 versus 0.0389 for SWVA ratings and PGR ratings in the $k = 3$ model, respectively).

Table C.1

Analysis Results of the Dominance of IP, PGR, SLO, and SWVA Measures in the Prediction of PPfT Final Rating

Base model	Base model R^2	Additional R^2 contribution of:			
		IP	PGR	SLO	SWVA
R^2 for individual components (conditional dominance *k* = 0)		0.52	0.4	0.3	0.15
IP	0.52		0.07	0.18	0.13
PGR	0.40	0.19		0.20	0.10
SLO	0.30	0.40	0.30		0.09
SWVA	0.15	0.50	0.35	0.24	
Average additional R^2 for adding one component to a one-component model (conditional dominance *k* = 1)		0.36	0.24	0.21	0.11
IP, PGR	0.59			0.16	0.11
IP, SLO	0.70		0.05		0.09
IP, SWVA	0.65		0.05	0.14	
PGR, SLO	0.60	0.15			0.06
PGR, SWVA	0.49	0.20		0.17	
SLO, SWVA	0.39	0.40	0.27		
Average additional R^2 for adding one component to a two-component model (conditional dominance *k* = 2)		0.25	0.12	0.16	0.09
IP, PGR, SLO	0.75				0.07
IP, PGR, SWVA	0.70			0.13	
IP, SLO, SWVA	0.79		0.04		
PGR, SLO, SWVA	0.66	0.16			
Average additional R^2 for adding one component to a three-component model (conditional dominance *k* = 3)		0.16	0.04	0.13	0.07
IP, PGR, SLO, SWVA	0.83				
Overall average additional R^2 for adding one component to all *k* models (general dominance)		0.32	0.20	0.20	0.10

Source. PPfT appraisal results from 2017–2018 and SAS EVAAS teacher-level value-added scores for 2017–2018.
Note. The sample was limited to n = 4,515 teachers on the standard PPfT appraisal plan. All values were rounded to 2 decimal places.

Appendix D

Tables D.1 and D.2 show results of factor analysis on the four components of the PPfT appraisal. The Kaiser-Meyer-Olkin (KMO) statistic was greater than 0.5, indicating that the proportion of variance in the four components could be caused by an underlying factor structure. Bartlett's Test of Sphericity was significant, indicating the correlation matrix is not an identity matrix, and therefore factor analysis may be suitable (Table D.1). A two-factor solution resulted from factor analysis on the four components of the PPfT appraisal (Table D.2). IP and PGR loaded into the first factor. SLOs and SWVA loaded into the second factor.

Table D.1
Kaiser-Meyer-Olkin Statistic for the Current Four Appraisal Components: IP, PGR, SLO, and SWVA

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.562
Bartlett's Test of Sphericity	approx. chi-square	895.495
	df	6
	sig.	0.000

Source. PPfT appraisal results from 2017–2018.

Table D.2
Pattern Matrix for the Current Four Appraisal Components: IP, PGR, SLO, and SWVA

	Factor	
	1	2
IP	0.934	
PGR	0.611	
SLO		0.350
SWVA		0.689

Source. PPfT appraisal results from 2017–2018.
Note. The extraction method was Principal Axis Factoring. The rotation method was Promax.

Tables D.3, D.4, and D.5 show the results of factor analysis on the hypothetical group of components inclusive of teacher value-added ratings with IP, PGR, and SLOs. Table D.3 shows the improved correlation matrix when teacher value-added correlations were greater or better than SLOs with IP and PGR. Table D.4 shows the improved internal consistency achieved by replacing SWVA ratings with teacher value-added ratings. For the factor analysis, the KMO statistic was greater than 0.5, indicating that the proportion of variance in the four components could be caused by an underlying factor structure. Bartlett's Test of Sphericity was significant, indicating the correlation matrix is not an identity matrix and therefore factor analysis may be suitable (Table D.5). A one-factor solution resulted from factor analysis on the hypothetical group of components inclusive of teacher value-added with IP, PGR, and SLOs (Table D.6).

Table D.3

Correlation Matrix for the Hypothetical Group of Components Inclusive of Teacher Value-Added Ratings With IP, PGR, and SLOs

	Pearson correlation coefficients and p values (*n* = 1,538)			
	IP rating	PGR rating	SLO rating	SWVA rating
IP rating	1	0.59 $p < 0.001$	0.24 $p < 0.001$	0.23 $p < 0.001$
PGR rating	0.59 $p < 0.001$	1	0.18 $p < 0.001$	0.23 $p < 0.001$
SLO rating	0.24 $p < 0.001$	0.18 $p < 0.001$	1	0.26 $p < 0.001$
Teacher value-added rating	0.23 $p < 0.001$	0.23 $p < 0.001$	0.26 $p < 0.001$	1

Source. PPfT appraisal results from 2017–2018 and SAS EVAAS teacher-level value-added scores for 2017–2018.

Table D.4

Standardized Cronbach's Alpha Coefficients for the Hypothetical Group of Components Inclusive of Teacher Value-Added Ratings With IP, PGR, and SLOs

Overall standardized Cronbach's alpha coefficient	Deleted strand	Adjusted standardized Cronbach's alpha coefficient with deletion
0.620	IP rating	0.467
	PGR rating	0.495
	SLO rating	0.619
	Teacher value-added rating	0.604

Source. PPfT appraisal results from 2017–2018 and SAS EVAAS teacher-level value-added scores for 2017–2018.

Table D.5

Kaiser-Meyer-Olkin Statistic for the Hypothetical Group of Components Inclusive of Teacher Value-Added Ratings With IP, PGR, and SLOs

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.608
Bartlett's Test of Sphericity	approx. chi-square	928.445
	df	6
	sig.	0.000

Source. PPfT appraisal results from 2017–2018 and SAS EVAAS teacher-level value-added scores for 2017–2018.

Table D.6

Factor matrix for the hypothetical group of components inclusive of teacher value-added with IP, PGR, and SLOs.

	Factor
	1
IP	0.783
PGR	0.711
SLO	0.331
Teacher value-added rating	0.354

Source. PPfT appraisal results from 2017–2018.
Note. Note. The extraction method was Principal Axis Factoring. Although Promax rotation was selected, no rotation was performed, given the one-factor solution.

References

Azen, R., & Budescu, D. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods, 8(*2), 129–148.

Hutchins, S. D. (2019). *Measurement validity and reliability of Professional Pathways for Teachers ratings* (DRE Publication 18.17 RB). Austin, TX: Austin Independent School District.

Hutchins, S. D., Looby, K., DeBaylo, P. & Leung, J. (2019). *Professional Pathways for Teachers: Summary of evaluation findings from 2017–2018* (DRE Publication 17.60). Austin, TX: Austin Independent School District.

AUSTIN INDEPENDENT SCHOOL DISTRICT

Author
Shaun D. Hutchins, Ph.D.

Department of Research and Evaluation
1111 West 6th Street, Suite D-350 | Austin, TX 78703-5338
512.414.1724 | fax: 512.414.1707
www.austinisd.org/dre | Twitter: @AISD_DRE

June 2019
Publication 18.17