Shaun D. Hutchins, Ph.D.

Publication 18.17 RB
June 2019

Measurement Validity and Reliability of Professional Pathways for Teachers

Research Brief

PPfT

Overview

The purpose of this Professional Pathways for Teachers (PPfT) evaluation research brief was to summarize findings from measurement validity and reliability analyses of PPfT appraisal data from the 2017–2018 school year. Detailed methods and results are available in the full technical report, DRE Publication 18.17 (Hutchins, 2019). The validity and reliability analyses were conducted in response to question from the PPfT oversight committee, district leadership, and program staff.

PPfT is a human capital system that blends four primary components: teacher appraisal, teacher professional development (PD) opportunities, teacher leadership opportunities, and teacher compensation. The PPfT appraisal component 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. PPfT appraisal yields an annual summative score from the measures of teaching quality that results in one of 5 possible final ratings for teachers: distinguished, highly effective, effective, minimally effective, and ineffective.

The validity and reliability of PPfT appraisal related to two basic ideas: did we measure what we intended to measure, and can we measure it consistently? That is, how well did the appraisal system measure teaching quality and how consistently was teaching quality measured? Validity and reliability analyses examined the psychometric properties of the PPfT appraisal instrument. 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.

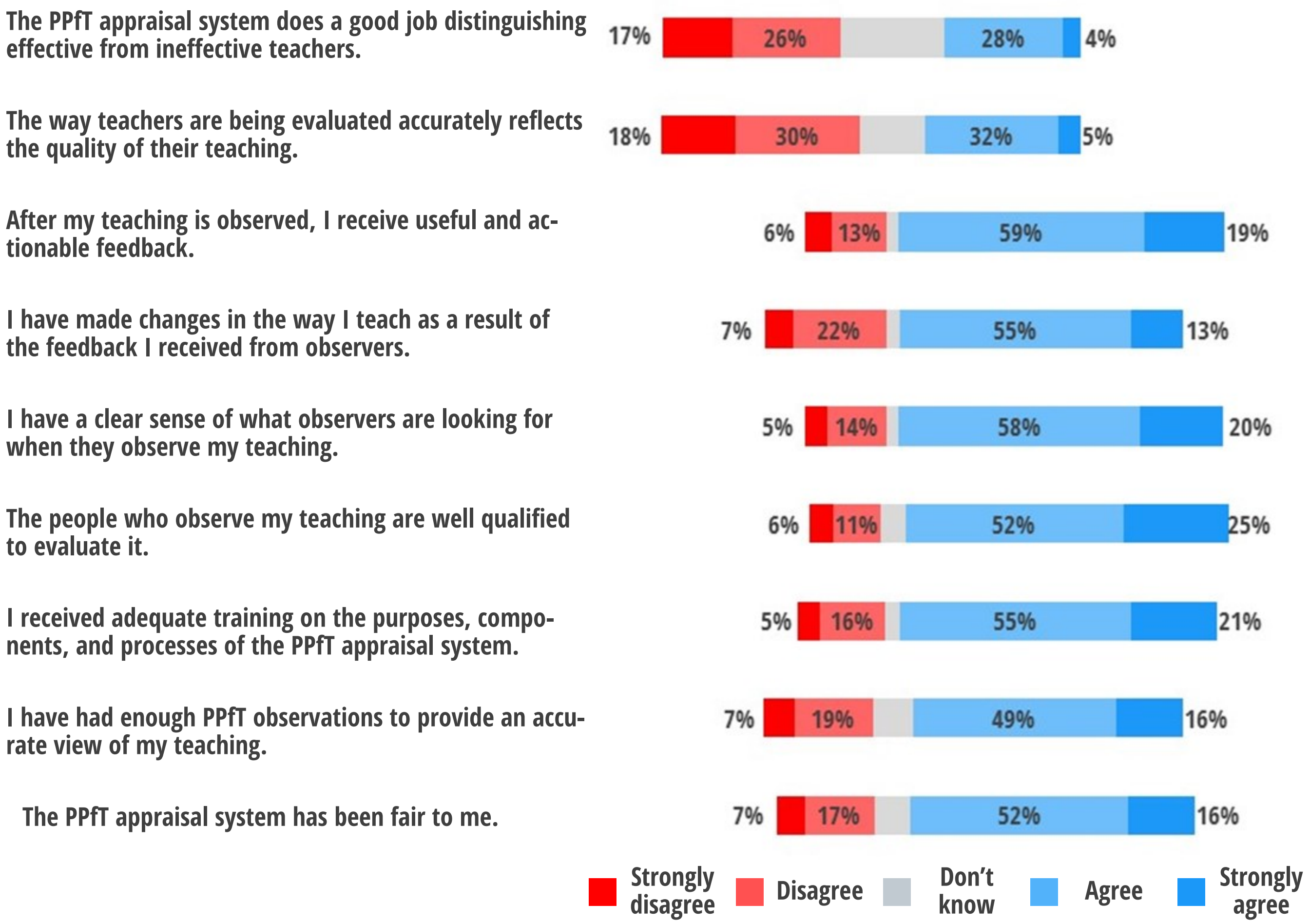
Content Validity: Did stakeholders feel the final and instructional practice ratings reflected the quality of their teaching?

Content validity analyses examined stakeholder perceptions about PPfT gathered from the spring 2018 AISD Employee Coordinated Survey (ECS). The items analyzed for content validity asked stakeholders whether they felt their 2017–2018 PPfT final ratings and 2017–2018 PPfT IP ratings reflected the quality of their teaching. Analyses of 2017 PPfT ECS items suggested strong content validity round the entire instructional practice process. However, stakeholders seemed divided on their perceptions of

Department of Research and Evaluation
Austin Independent School District

whether the appraisal system measures teaching quality. Potential issues with item design (i.e., first person versus third person) and need for more education on PPfT were considered (Figure 1).

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.

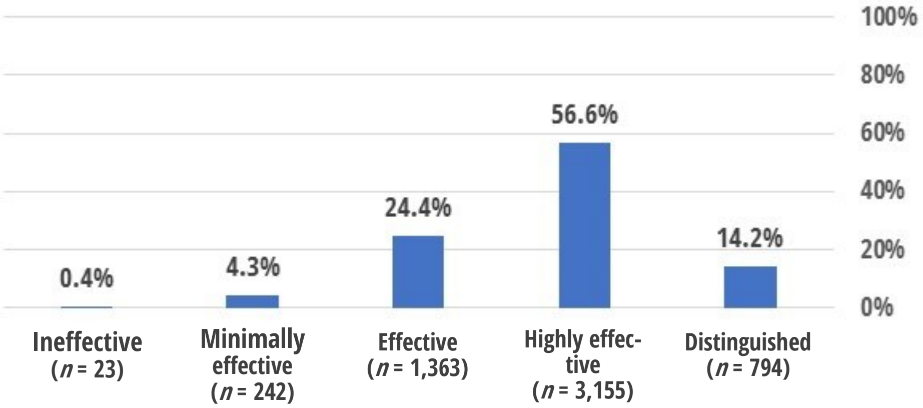


Source. 2017–2018 Employee Coordinated Survey.
Note. Of the 5,577 teachers appraised under PPfT in 2017–2018, approximately 8% (*n* = 470) responded to the survey.

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

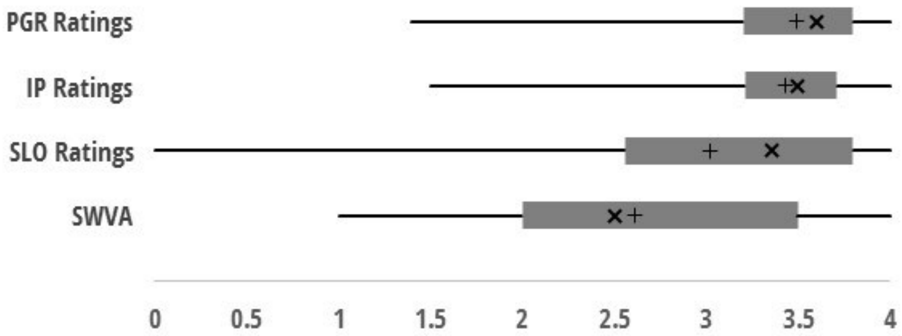
Concurrent validity analyses examined the distributions of 2017–2018 PPfT final ratings and PPfT appraisal component scores. Each scale was analyzed to assess the extent that teachers were differentiated in the distributions of measures. The overall differentiation of teaches across final rating categories (i.e., distinguished, highly effective, effective, minimally effective, and ineffective) suggests inter-category concurrent validity (Figure 2). However, the large mode of teachers receiving a highly effective final rating (i.e., 56.6%) suggests weaker intra-category concurrent validity. Analysis of PPfT appraisal components shows differentiation between teachers at the component-level was best for SWVA and SLOs, worst for IP and PGR (Figure 3). A shift in the procedures for rater calibration around differentiating 2s, 3s, and 4s on the instructional practice rubric was considered as means to maintain inter-category concurrently validity while potentially improving intra-category concurrent validity.

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.

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?

Convergent validity analyses examined the relationship between teachers' final ratings and their students' growth. Correlation analyses between 2017–2018 PPfT final ratings and 2017–2018 SAS EVAAS teacher value-added data assessed if teaching quality was associated with student growth. Associations between 2017–2018 PPfT final ratings and 2017–2018 SAS EVAAS teacher value-added data 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. The exception was in grade 8 and on the U.S. history end-of-course (EOC) assessments where findings yielded mixed results (Table 1).

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

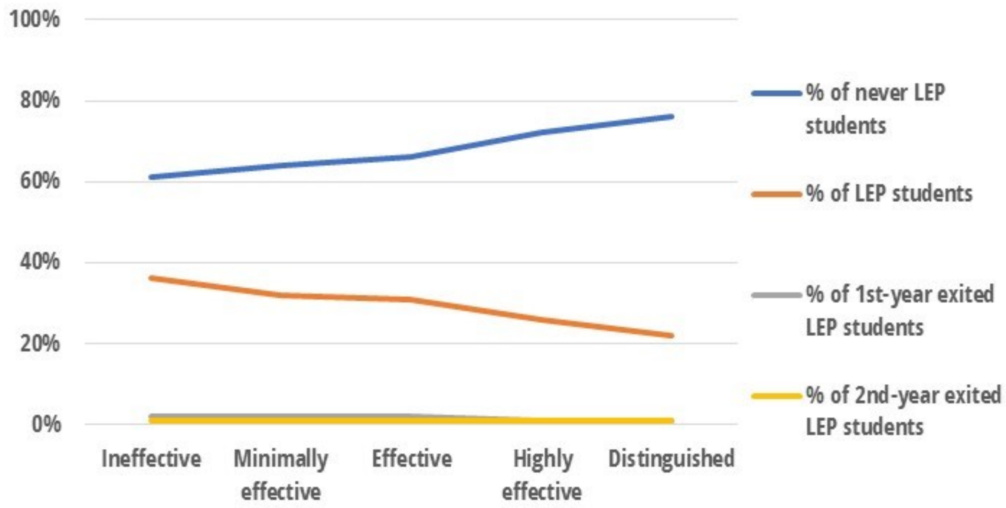
Tested subject	Tested grades					
	4	5	6	7	8	Secondary (EOC)
Math	+	+	+	+	NR	NA
Reading	+	+	+	+	NR	NA
Science	NA	+	NA	NA	NR	NA
Writing	NA	NA	NA	+	NA	NA
Social studies	NA	NA	NA	NA	+/NR	NA
Algebra I	NA	NA	NA	NA	NA	+
Biology	NA	NA	NA	NA	NA	+
English I	NA	NA	NA	NA	NA	+
English II	NA	NA	NA	NA	NA	+
U.S. History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

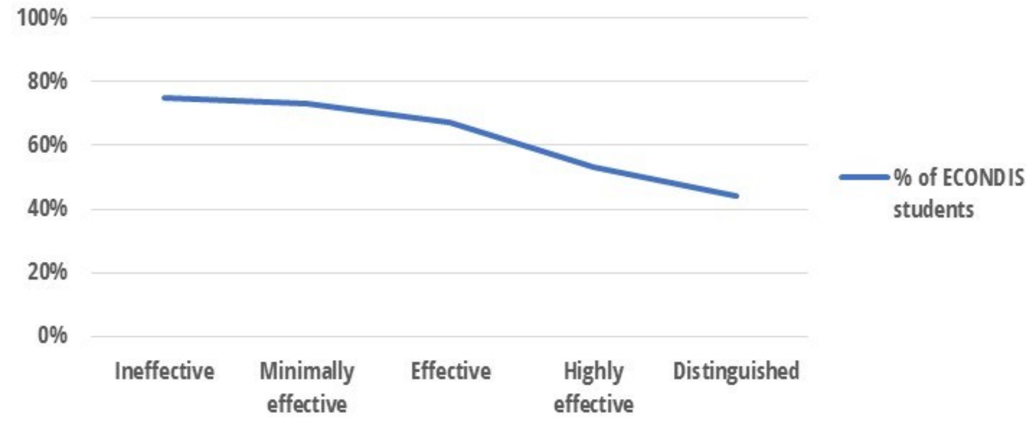
Discriminant validity analyses examined the relationship between teachers' final ratings and their students' demographic characteristics. Correlation analyses between 2017–2018 PPfT final ratings and characteristics of the students served by teachers assessed if ratings of teaching quality were independent of the characteristics of the students taught by the teachers. Associations between 2017–2018 PPfT final ratings and their students' demographic characteristics 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 (Figure 4), economically disadvantaged (ECONDIS) status (Figure 5), and the race/ethnicity (Figure 6) 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 considered as a potential lever to equitably distribute high-quality teachers with populations of underserved students.

Figure 4.
Student LEP status was associated with teachers' PPfT final ratings.



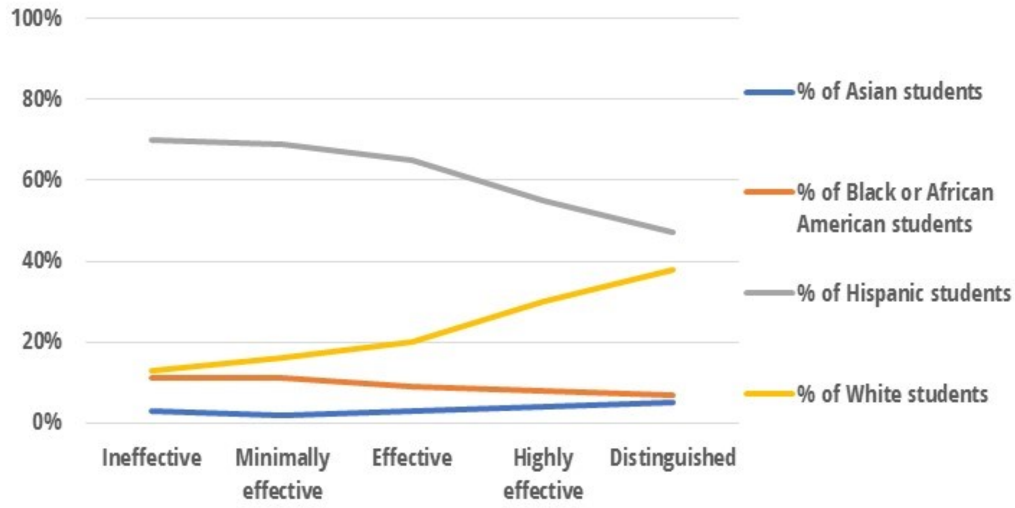
Source. PPfT appraisal results from 2017–2018 and student demographic information from 2017–2018 Texas Student Data System (TSDS).

Figure 5.
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).

Figure 6.
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.

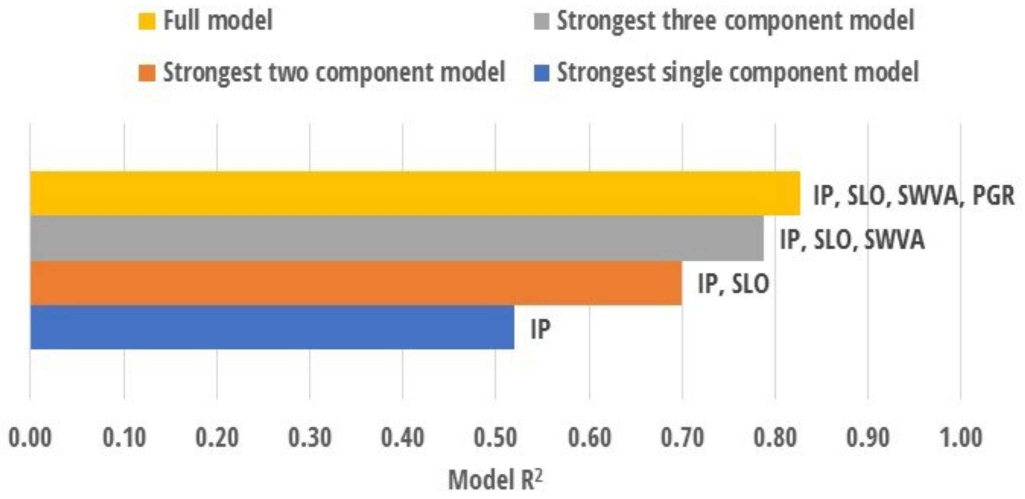


Source. PPfT appraisal results from 2017–2018 and student demographic information from 2017–2018 Texas Student Data System (TSDS).

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

Dominance analyses examined relative importance of appraisal components for predicting teachers' final ratings. Multiple regression analyses predicting 2017–2018 PPfT final ratings were conducted to examine the additional R^2 contribution of each component in models of all possible combinations of PPfT appraisal components. 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 suggested 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. Adjustments to rater calibration and rigor should be considered in parallel with work to improve intra-category concurrent validity (Figure 7).

Figure 7.
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 all components of PPfT appraisal, only IP was rated by two different raters. Consequently, interrater reliability analyses examined the degree of agreement between raters who scored teachers' instructional practices. T-tests, correlations, and agreement (i.e., Cohen's weighted kappa coefficient) were run between fall and spring ratings on all 7 strands of 2017–2018 PPfT IP ratings. Interrater reliability analyses were inconclusive due to confounds between raters and time and teacher improvement. 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) and teachers use their fall observation feedback to improve their craft for their subsequent spring observation. Despite the confounds, the collective set of analyses on interrater reliability pointed towards adequate agreement between raters. The use of floating peer observers to partner with school administrators during both observations was considered as means to reduce confounds in analysis of interrater reliability.

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

Internal consistency analyses examined the appraisal strands (i.e., Fall IP ratings, spring IP ratings, and PGR ratings) and appraisal component scores (i.e., IP, PGR, SLOs, and SWVA) for consistent patterns of scoring. Cronbach's alpha was used to examine the consistency of scores within the 7 strands of 2017–2018 PPfT IP, the 5 strands of 2017–2018 PPfT PGR, and the 4 components of 2017–2018 PPfT final ratings. Fall IP ratings, spring IP ratings, and PGR ratings all showed evidence of strong internal consistency (i.e., Cronbach's alpha ranged from 0.81 to 0.87 internal consistency could not be improved by removing any strands). However, the set of four appraisal components (i.e., IP, PGR, SLOs, and SWVA) showed evidence of somewhat weaker internal consistency (Table 2). 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 (Table 3) and change the factor analytic structure from a two-factor solution (IP and PGR in one factor and SLOs and SWVA in the second) to a single factor solution comprising IP, PGR, SLOs, and teacher value-added.

Table 2.
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.

Table 3.
The standardized Cronbach's Alpha Coefficient improves using a hypothetical group of appraisal 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.

Summary of Findings

How well did the appraisal system measure teaching quality in 2017–2018? Validity analyses generally showed evidence for content, concurrent, and convergent validity of PPfT appraisal, jointly suggesting valid measurement of quality teaching by the appraisal instrument. Discriminant validity findings were mixed, showing that the gender, GT status, and SPED status of the students served by teachers appeared to operate independently of the final ratings teachers received. However, the LEP status, ECONDIS status, and the race/ethnicity (percentage Hispanic and White only) of the students served by teachers appeared to operate in some dependency with the final ratings teachers received. The inclusive discriminant validity results show that the final ratings received by teachers operated independently of some, but not all of their students' characteristics. Results of dominance analyses highlighted the importance of actively working to avoid ceiling effects in any of the rating scales. In some ways, the appraisal instrument is as accurate as it is applied. When the distributions of rated values given to teachers start to cluster at the high end of the scale, as shown with PGR and IP ratings in Figure 3, the scale begins to lose its capacity to adequate differentiate teachers on the intended teaching quality construct.

How consistently was teaching quality measured? Reliability analyses generally suggested consistent measurement of teaching quality, particularly among the campus administrator rated parts of PPfT appraisal. The limited range of ratings on IP and PGR provided may have been something that factored into the consistency of those two components (Figure 3). Agreement between raters on IP seemed adequate, but confounds between raters and time and within year teacher improvement precluded conclusive assessment of rater agreement. The juxtaposition of strong internal consistency of campus administrator rated items with the adequate, yet weaker internal consistency of the four appraisal components highlights the need for ongoing discussion and explicit valuing around collective and individual attribution to student growth measurement (e.g., our students versus my students). Comparisons of the internal consistency and factor analytic solutions with school-wide value-added and teach value-add should provide data for these conversations.

References

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

AUSTIN INDEPENDENT SCHOOL DISTRICT

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 RB