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).

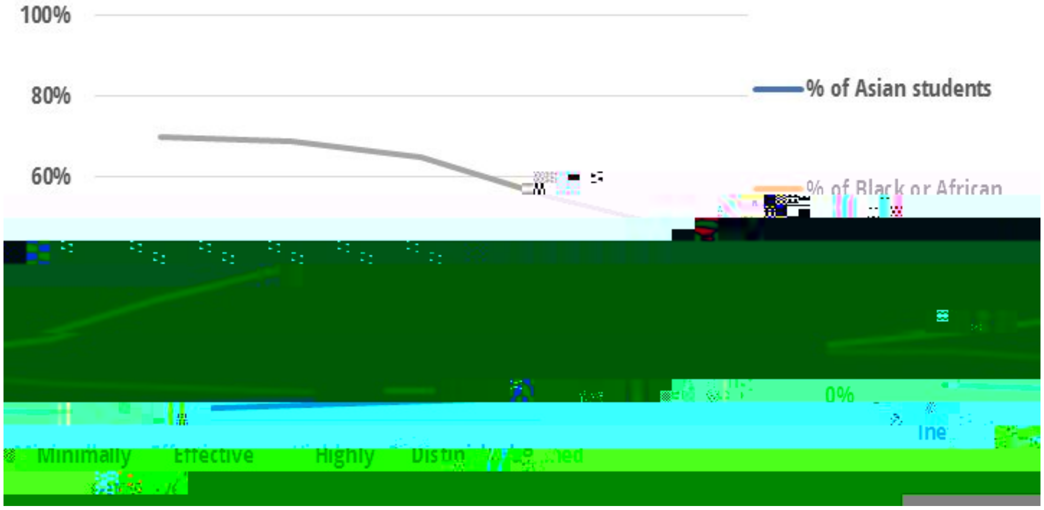
9GD7
FG67@F EF3FGE I 3E 3EEA5 3F76 I F: F735: 7DE 8 @3 D3F @9E

AGD57 8 3BBD3 E3 D7EG FE 8DA 3@6 EFG67@F 67 A9D3B: 5 @8AD 3F A@8DA 7J3E FG67@F 3F3 EF7

9GD7
FG67@F EF3FGE I 3E 3EEA5 3F76 I F: F735: 7DE 8 @3 D3F @9E

AGD57 8 3BBD3 E3 D7EG FE 8DA 3@6 EFG67@F 67 A9D3B: 5 @8AD 3F A@8DA 7J3E FG67@F 3F3 EF7

9GD7
FG67@F :7 B7D57@F397E A8) : F7 3@6 EB3@5 EFG67@FE E7DH76 I 7D7 3EEA5 3F76 I F: F735: 7DE 8 @3 D3F @9E 4GF F: 7 B7D57@F397E A8 E 3@3@6 8D53@ 7D53@I 7D7 @AF



AGD57 8 3BBD3 E3 D7EG FE 8DA 3@6 EFG67@F 67 A9D3B: 5 @8AD 3F A@8DA 7J3E FG67@F 3F3 EF7

Internal consistency analyses examined the appraisal strands (i.e., Fall IP ratings, spring IP ratings, and PGR ratings)

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