COMPARATIVE GROWTH MODEL ANALYSIS

Introduction

In May 2011, the HISD Board of Education adopted a new teacher appraisal and development system. Comparative Growth is included as one of the five measures in the Student Performance component of the system. It will most commonly be applied as a second Student Performance measure for teachers who have value-added growth, because under the system's guiding principles, a teacher must be evaluated on multiple measures, and never on value-added growth alone. It is also applied as the primary Student Performance measure for second-grade teachers of the core foundation subjects of English, Math and Language Arts.

Comparative Growth, like all Student Performance measures, will be applied for teacher appraisal purposes for the first time during the 2012-13 school year, based on spring 2013 data from the Stanford/Aprenda and TELPAS-Reading assessments. Because Comparative Growth is a new district measure to be used in teachers' appraisals, HISD undertook an additional evaluation of the Comparative Growth model in April 2012, using data from the 2010-11 school year. The analysis involved various research questions aimed at addressing additional questions that have been raised by stakeholders since the measure was adopted in April of 2011. By thoroughly vetting this model, the district is safeguarding against adverse effects the new measure might have on teachers' appraisal ratings.

The purpose of this document is to explain the results of this analysis. The research questions covered in this analysis were:

1. Reducing the minimum cohort/comparison group size of 25 would allow more students to be included in the analysis, but would doing so compromise the statistical rigor of the model and/or adversely affect Comparative Growth scores?

2. What is the correlation between Comparative Growth Teacher Performance Levels and EVAAS® Teacher Performance Levels for each subject and grade level where both measures are applied? Is that level of correlation appropriate given the different methodologies and assessments utilized by both measures?

3. Does the Comparative Growth model show adverse effects on the performance levels of teachers with large numbers of high-performing students, as measured by those students' year one Stanford/Aprenda scores? What adjustments, if any, should be made to ensure the measure is fairly applied to teachers?

4. The Student Performance Working Group recommended using Comparative Growth on the TELPAS assessment for English Language Learners (ELL) students in grades 3-12. However, many cohort groups in grades 9-12 do not meet the minimum *n* size (25 students) for the model. Therefore, they are excluded from the analysis. Given the small number of students with TELPAS scores in high schools, should TELPAS be used as a measure for high school teachers?

5. Are the Comparative Growth cut scores for the four Teacher Performance Levels (1, 2, 3, 4) appropriate for all subjects and grade levels? Are teachers in certain grades/school

© 2012, Houston Independent School District. All Rights Reserved.

levels unduly advantaged or disadvantaged by these cut scores and should adjustments be made to ensure fairness?

Analysis of three of these five questions yielded conclusions that led the research team to recommend adjustments to the initial Comparative Growth Model and yearly analysis to monitor outcomes. The remainder of this paper details the Comparative Growth Model; discusses each of the five research questions and conclusions in this analysis; and summarizes the recommended adjustments to the application of the Comparative Growth measure, which were accepted for use beginning in the 2011-12 school year.

Comparative Growth Model

Comparative Growth measures the progress of a teacher's students on a given assessment compared to all other students within the same school district who start at the same test-score level. The Comparative Growth model has its origins in the principles of the Colorado Growth Model and the work of Barlevy & Neil. As applied in HISD, Comparative Growth is based on the Stanford/Aprenda in certain subjects in grades 2-8, and on TELPAS-Reading scale scores in grades 3-8.

Calculating Comparative Growth involves a number of steps and processes, which are described below. For a more visual explanation of this model, see *Student Performance Training Session #3: Comparative Growth*. This and additional resources are also available on the ASPIRE Portal under the "CG Reports" section of the My ASPIRE dashboard.

Comparative Growth Method for Stanford/Aprenda
The process of calculating the Comparative Growth Teacher Median for Stanford and Aprenda is as follows:

A. For each subject and grade level of the assessment, students are grouped by the language of the tests they took. This categorization process yields three groups of students for each subject and grade level of the test. Those who took:
   - Stanford in the previous year and Stanford in the current year
   - Aprenda in the previous year and Aprenda in the current year
   - Aprenda in the previous year and Stanford in the current year

B. After being placed in groups based on test language over two years, students are placed in sub-groups (called comparison groups) based on their prior year's testing performance. For example, all students who took Stanford both years and received a Normal Curve Equivalent score (NCE) of 52 on the previous year's test will be placed in the same comparison group. Prior-year NCE is considered the student's starting point, and students are only compared against other students in the district with the same starting point.

C. Within comparison groups, students are percentile-ranked within HISD using the current year's test scores. This district percentile-rank is the student's growth percentile.

D. Finally, a teacher's Comparative Growth Teacher Median is calculated by taking the Median Growth Percentile Score of the students in his/her class. Appraisers translate the teacher's Median Growth Score into the Teacher's Performance Rating for his/her appraisal using conversion tables (below).

© 2012, Houston Independent School District. All Rights Reserved.

Comparative Growth Method for TELPAS
The process of calculating the Comparative Growth Teacher Median for the TELPAS-Reading assessment for English Language Learners (ELLs) in grades 3-8 is similar as for Stanford and Aprenda. However, rather than using NCEs or the state English language proficiency levels (Beginning, Intermediate, Advanced, Advanced High), **scale scores** are used because they allow teachers to show growth with students *within* proficiency levels. For example, a student might score at the Intermediate proficiency level two years in a row, but in reality did acquire more English, which comparing the scale score from one year to the next can show.

Only the reading portion of the TELPAS assessment is used because: 1) It is weighted more heavily (75 percent) than the other domains of the test and 2) Scores for Reading are represented by a vertical scale score derived from an objective multiple-choice assessment. For Comparative Growth, district-wide comparison groups are formed based on prior-year scale score on the TELPAS-Reading assessment. All students with the same scale score the previous year form one comparison group and are percentile-ranked based on the current year's scale score.

Scoring Comparative Growth
Below are the tables for converting Teacher Medians to Performance Levels:

Stanford/Aprenda, Grades 2-8

Elementary Performance Levels

Comparative Growth Elementary Teacher Median	Comparative Growth Performance Level
<28	1
28-47	2
48-68	3
>68	4

Secondary Performance Levels

Comparative Growth Secondary Teacher Median	Comparative Growth Performance Level
<33	1
33-49	2
50-64	3
>64	4

TELPAS-Reading, Grades 3-8

Comparative Growth Teacher Median on TELPAS (Gr. 3-8)	Comparative Growth Performance Level
<28	1
28-46	2
47-66	3
67+	4

© 2012, Houston Independent School District. All Rights Reserved.

Special Situations
There are some situations where teachers who would otherwise receive a Comparative Growth score may not receive one. In certain instances, students will be excluded from Comparative Growth calculations to ensure an equal advantage to all teachers.

Situations where teachers will not receive Comparative Growth scores include the following:
1. Teachers who have fewer than seven students per subject and grade level linked 30 percent or more of the school year to their rosters. These teachers will not have enough students with Student Growth Percentile scores to calculate a meaningful teacher Comparative Growth Rating.
2. Teachers whose class composition is greater than 40 percent students identified as special education. The model cannot reliably calculate Comparative Growth Teacher Medians for these teachers.

Situations where a student will be excluded from Comparative Growth calculations include students who:
1. Are missing one of the two required test scores. This includes students who may be new to the district, state, or country, take TELPAS, Stanford, or Aprenda for the first time, and therefore have no prior year score.
2. Do not follow a traditional grade progression in two consecutive years (e.g., skipped grades or were not allowed to progress to the next grade level). [1]
3. Fall into district-wide comparison groups with fewer than 25 students. This is because groups smaller than 25 are not large enough to have a broad distribution to calculate student growth percentile scores.
4. Are linked less than 30 percent of the school year to a teacher's roster. Teachers do not have enough time with these students to influence their scores substantially.

Research and Analysis

Research Question 1
Reducing the minimum cohort/comparison group size of 25 would allow us to include more students in the analysis, but would doing so compromise the statistical rigor of the model and/or adversely affect Comparative Growth scores?

The Comparative Growth model was designed using student cohorts of 25 students or more to ensure that the comparison group was large enough to accurately assess student growth. A side effect of this decision is that a Student Growth Percentile cannot be calculated for approximately 6 percent of students taking Stanford/Aprenda because their comparison group is not large enough.[2]

Because a teacher must have at least seven students per subject and grade level with Student Growth Percentiles to receive a Comparative Growth Performance Level, an analysis was conducted to see how many teachers would be affected by excluding students whose comparison groups did not meet the minimum *n* size of 25. The majority of teachers' performance levels were *not* affected by the students who were excluded from the Comparative Growth calculations. Additionally, because Comparative Growth is calculated using actual students instead of a theoretical curve, teachers could potentially be disadvantaged if cohorts smaller than 25 were included.

[1] Constraint only applies to Stanford/Aprenda, not TELPAS
[2] Constraints regarding student cohorts for TELPAS are further discussed in Research Question 4.

© 2012, Houston Independent School District. All Rights Reserved.

Additional analysis was conducted to examine the impact of a teacher having the greater number of excluded students who were clustered in the lower grades as a result of students participating in early-exit bilingual programs. Comparative Growth calculations control for Spanish-to-English transition because HISD measures the progress of these students against other students who are also transitioning to English. Students are compared only to other students in the district who took the same tests as they did two years in a row (e.g., those who took Aprenda one year and Stanford the next). Because there are fewer students in early-exit programs, these students were more likely to be in a comparison group with fewer than 25 students. While the analysis showed that students participating in an early-exit bilingual program did not disadvantage teachers, it was further determined that no teachers would be disadvantaged as the program is being discontinued.

Conclusion: The comparison cohort size should remain at 25 students or more to maintain the integrity of the model.

Research Question 2

What is the correlation between Comparative Growth Teacher Performance Levels and EVAAS® Teacher Performance Levels for each subject and grade level where both measures are applied? Is that level of correlation appropriate given the different methodologies and assessments utilized by both measures?

While one of the hallmarks of the student performance component of the teacher appraisal and development system is that every teacher will have multiple measures to provide a more complete picture of true teacher performance, there is also an implicit understanding that for most teachers, the resulting performance levels from their various measures will be somewhat aligned. Outlying cases might warrant further investigation, and so the analysis sought to determine the extent of misalignment (i.e., for how many teachers who have both value-added and Comparative Growth might the performance levels on those two measures be misaligned).

As such, it was important to look at the relationship between EVAAS® Teacher Performance Levels and Comparative Growth Performance Levels. The two measures were compared for teachers who could have both measures; see appendix for data table. Once the performance levels were compared, it was noted that 67, or 3 percent of all 2,536 teacher reports, would have combinations where there is a very low EVAAS® Performance Level (-2 or -1) and a very high Comparative Growth Performance Level (4), or a very high EVAAS® performance level (+2 or +1) and a very low Comparative Growth Performance Level (1).

Conclusion: The current level of correlation is appropriate given that the measures are based on different student assessments and use different statistical methodologies, and therefore should give a balanced view of a teacher's performance when used together.

Research Question 3

Does the Comparative Growth model show adverse effects on the performance levels of teachers with large numbers of high-performing students, as measured by those students' year one Stanford/Aprenda scores? What adjustments, if any, should be made to ensure measure is fairly applied to teachers?

© 2012, Houston Independent School District. All Rights Reserved.

The Department of Research and Accountability had determined that teachers of students with NCE scores of 99.0 for the two consecutive years would have their Comparative Growth Percentile Rank set at 99.0 so that no teacher would be disadvantaged by having students scoring at the highest level of the test. Analysts first set out to understand how substantial a difference in learning was represented by the change from a student's NCE score of 99.0 in year 1 to an NCE score of 93.3 or 89.6 in year 2. (Descending from an NCE score of 99.0, the next two possible scores are an NCE of 93.3 and an NCE of 89.6.) After examining the technical manuals for Stanford/Aprenda and reviewing the learning represented by each score, it was determined that a student who receives an NCE score of 99.0 in one year and an NCE score of 89.6 in the next year has scored substantively differently on the tests in those two years. However, a student who receives an NCE score of 99.0 in one year and an NCE score of 93.3 in the next year may not have learned a substantively different amount in that second school year. As such, the effect of students moving from NCE scores of 99.0 to 93.3 warranted further exploration, while those moving from NCE scores of 99.0 to 89.6 did not.

Analysts examined students who had an NCE score of 99.0 in the first year of testing and an NCE score of 93.3 in the second year of testing using student data from the 2009-2010 and 2010-2011 school years, respectively. It was determined that 736 students, assigned to 578 classes, taught by 496 teachers, fell into this situation. The analysts looked at the effect of this change on both the median scores for the students and the teachers' final ratings. After running the simulation, it was shown that while the Median Score for each student had the potential to change significantly, it did not typically affect the appraisal rating (90 percent of the classes saw no change in the appraisal rating). Yet, 10 percent of classes whose teachers would have had a rating change warranted additional analysis. Further analysis showed that for those students who scored 99.0 NCE in year 1 and 93.3 in year 2, locking those students' year 2 percentile ranking at 98, which is the corresponding percentile rank for 93.3 NCE, did not disadvantage their teachers in the Comparative Growth model.

Conclusion: To prevent teachers from being unfairly disadvantaged when their students are at the 99th percentile in year one scores and at the 99th or 93rd percentile for year two scores, they are given a district percentile rank of 99 and 98 respectively in the Comparative Growth model.

Research Question 4

The Student Performance Working Group recommended using Comparative Growth on the TELPAS assessment for English Language Learners (ELL) students in grades 3-12. However, many cohort groups in grades 9-12 do not meet the minimum n size (25 students) for the model. Therefore, they are excluded from the analysis. Given the small number of students with TELPAS scores in high schools, should TELPAS be used as a measure for high school teachers?

Another focus of the analysis is the number of ELL students taking TELPAS in grades 9-12 who are excluded from the model because their cohort does not have 25 students. Given the previous considerations around cohort size, lowering the cohort requirements for this group of students was not considered as an option. The subsequent consideration was to create cohorts of students across grade levels given that the TELPAS exam is the same for grades 8-9 and grades 10-12 (e.g., combine 10th, 11th, and 12th grade students with the same year 1 scale score into one comparison group). This option would mean that students would be considered alongside students in different grades who had the same score on TELPAS that school year. This change would allow more teachers to use Comparative Growth.

© 2012, Houston Independent School District. All Rights Reserved.

Analysts investigated what impact would occur because while students may have been taking the same test, students who were in different grades actually had reasonably different performance expectations based on how many years they had taken the specific TELPAS exam. For example, a 10th grade student taking the grades 10-12 TELPAS exam for the first time who scores the same as an 11th grade student taking the test for the second time may not be similar enough academically to be considered in the same cohort. In other words, although TELPAS is a language proficiency test, and language proficiency level does not correspond to grade level, the 11th grade student taking the grades 10-12 TELPAS assessment has one additional year of exposure to the academic language on the test. He and his teacher might be unfairly advantaged (and the 10th grade student and his teacher unfairly disadvantaged) in the Comparative Growth model. Without this option, there are not enough students taking TELPAS in grades 9-12 to provide appropriate comparison groups with which to calculate Comparative Growth.

Conclusion: There are not enough students to calculate Comparative Growth accurately and fairly for students using TELPAS in grades 9-12. Therefore, Comparative Growth will not be included as potential measure for teachers whose students use TELPAS in grades 9-12.

Research Question 5

Are the Comparative Growth cut scores for the four Teacher Performance Levels (1, 2, 3, 4) appropriate for all subjects and grade levels? Are teachers in certain grades/school levels unduly advantaged or disadvantaged by these cut scores and should adjustments be made to ensure fairness?

To ensure that the Comparative Growth model is set up to assess students in all courses equally, analysis was conducted to ensure that the cut points set for the four performance levels resulted in reasonably comparable distributions of performance levels for different grades and subjects. Using student performance data from the 2010-2011 school year, Comparative Growth Levels were calculated for all grades and subjects.

To determine what would constitute a disproportionate number of teachers, the analysis began looking at the Teacher Median of 50, which represents typical growth. Similar to a Student Growth Percentile of 30 indicating that 70 percent of peers achieving greater growth, a Teacher Median shows the relative amount of growth a teacher is able to help students achieve. Using this median as a starting point, data were examined for each test to ensure that cut scores appropriately recognized effective teaching, as measured by these assessments.

The analysis indicated that different cut scores were needed for elementary and secondary teachers in the Stanford/Aprenda model due to the differences in the expectations of the tests and the numbers of students assigned to each teacher of record. While there were no significant disparities among the grades and subjects within school levels, there were disparities between elementary and secondary teachers. As such, further analysis was conducted to determine the best cut scores for each.

Conclusion: When considering the goal of ensuring that disproportionate numbers of teachers did not receive particular performance levels and were not unfairly disadvantaged, there should be different cut scores set for elementary and secondary teachers in the Stanford/Aprenda model. See the cut scores determined from this analysis on page 3.

© 2012, Houston Independent School District. All Rights Reserved.

Recommendations and Actions Taken

In summary, the results to three of the research questions warranted adjustments to the initial Comparative Growth model. These address concerns related to teachers with large numbers of high performing students; the use of this model using TELPAS for high schools; and the cut scores for the teacher performance levels.

Recommendation regarding teachers with large numbers of high performing students: To prevent some teachers from being unfairly disadvantaged when their students are at the 99th percentile in year one scores and at the 99th or 93rd percentile for year two scores, they are given a district percentile rank of 99 and 98 respectively in the Comparative Growth model.

Recommendation regarding Comparative Growth on TELPAS for high school ELLs: Due to small comparison group sizes, TELPAS will not be used to generate Comparative Growth Scores for students in grades 9-12. Instead, Students' Progress on TELPAS-Reading should be used.

Recommendation regarding cut scores: The cut scores should be set differently for school levels (elementary and secondary) and test (Stanford/Aprenda and TELPAS). The recommended cut scores, which were based on the reading tests and then analyzed across all subjects to ensure consistency, can be seen on page 3.

The above recommendations have been incorporated into the 2011–2012 Comparative Growth model and will be reviewed annually.

© 2012, Houston Independent School District. All Rights Reserved.

Appendix

Research Question 2: *What is the correlation between Comparative Growth Teacher Performance Levels and EVAAS® Teacher Performance Levels for each subject and grade level where both measures are applied? What level of correlation is satisfactory?*

Comparison Between EVAAS® and Comparative Growth for All Grade Levels and Subjects			EVAAS® Performance Level					
			-2	**-1**	**NDD**	**1**	**2**	**ALL**
Comparative Growth Performance Level	**1**	Count	150	56	1	7	1	215
		%	52%	19%	0%	2%	0%	100%
	2	Count	146	141	16	31	16	350
		%	28%	27%	3%	6%	3%	100%
	3	Count	183	236	212	206	212	1,049
		%	11%	15%	13%	13%	13%	100%
	4	Count	19	40	355	153	355	922
		%	2%	5%	42%	18%	42%	100%

Based on 2,536 teacher reports in 2010–2011 by grade/subject generated for both EVAAS® and Comparative Growth

© 2012, Houston Independent School District. All Rights Reserved.