Comparing Through-Course and Cross-Year Summative Assessment Growth Scores

Introduction

The Race to the Top program has defined a through-course (TC) summative assessment as "a system component or set of assessment system components that is administered periodically during the academic year" (U. S. Department of Education, 2010, p. 18178). Through-course summative assessments aim to provide "an accurate measure of student achievement across the full performance continuum and an accurate measure of student growth over a full academic year or course" (U. S. Department of Education, 2010, p. 18178). Thus, TC summative assessments can be used for growth score calculations. Most growth scores, however, are calculated using cross-year (CY) summative assessment scores (Goldschmidt, Choi, & Beaudoin, 2012). Will growth scores calculated by the TC summative assessment be different from growth scores calculated by the CY summative assessment for same students? If there are differences, how big are the differences? Which assessment can provide more accurate growth scores? This study aimed to compare TC and CY growth scores to answer these questions.

The TC summative assessment can use one of the two designs: (1) Each TC summative assessment measures a segment of the curriculum, or (2) each TC summative assessment measures all the components of curriculum. The first design usually uses the multidimensional item response theory model. When this multidimensional model is used, maintaining the integrity of multiple dimension scales of different assessments is difficult, and combining different assessment results for the growth measurement is problematic (Mislevy & Zwick, 2012). The second design usually uses the unidimensional item response theory model. When this unidimensional model is used, the scores of different assessments can be put on the same scale easily. Since scores on the same scale are required by most growth models for growth

measurement, the TC summative assessment used in this study used the second design with the unidimensional item response theory model.

Method

Assessment

The TC summative assessment system used in this study included four assessments: three interim assessments and one end-of year assessment. Interim assessments were administrated three times in the 2011–2012 academic year, and the end-of year assessment was administrated at the end of the 2011–2012 academic year. These four TC summative assessments all measured the same curriculum standard, and their scores were all included in the calculation of TC growth scores. On the other hand, in order to compare two different growth scores, four end-of-year assessment scores were included in the calculation of CY growth scores. These end-of-year assessments were administrated at the end of the 2008–2009, 2009–2010, 2010–2011, and 2011–2012 academic years. The 2011–2012 end-of-year assessment was used in both growth score calculations. The four TC summative assessments were all grade 7 math assessments, while the four CY summative assessments were grade 4, grade 5, grade 6, and grade 7 math assessments. In all, there were seven assessments included in the study.

Linking

The scores of the four TC summative assessments were on the same scale. Thus, no linking was needed for the TC growth score calculation using these scores. However, the scores of the four CY assessments were not on the same scale because the vertical scaling was not used in this CY assessment system. To make the two growth scores more comparable, the four CY assessment scores should be on the same scale as the four TC summative assessment scores. Thus, the 2008–2009, 2009–2010, and 2010–2011 end-of-year summative assessment scores

were put on the score scale of the 2011–2012 end-of-year summative assessment. Since these assessments are cross-grade assessments without common items, the vertical scaling cannot be conducted. Thus, an equipercentile linking method with a single group design was applied (Kolen & Brennan, 2004) to allow for comparisons.

Sample

In this study, the author used two samples: one for calculating the growth models and one for linking the assessments. In all, 1,439 students in 39 different schools took all seven assessments. These 1,439 students, the sample for the growth model calculation, had scores for all assessments, so there was no missing data. For linking with the equipercentile linking method, a larger pool of data was used. A number of students, including the growth score sample, took the different end-of-year summative assessments in two years, and pairs of their scores were used to link the different pairs of assessments. The data are summarized in table 1.

Growth model

Mislevy and Zwick (2012) suggest using hierarchical growth modeling to calculate the growth scores of TC summative assessments. Thus, a two-level hierarchical linear model (HLM) was used for the growth score calculation (Raudenbush & Bryk, 2002).

The level 1 model was the within-student linear regression model, one for each student. Each linear equation regressed scale scores from four assessments on four time points. The first assessment was set at time zero, whereas the other three assessments were set at time one, two, and three, respectively. The level 1 model is:

$$Y_{ti} = \pi_{0i} + \pi_{1i}t + e_{ti}, \quad (1)$$

where Y_{ti} is the scale score for student i at time t, π_{0i} is the intercept of linear equation for student i, π_{1i} is the slope of the linear equation and is the growth rate for student i, t is the time point, and e_{ti} is the random error term.

The level 2 model included two regression equations modeling the intercepts and slopes of the level 1 regression functions. These two equations predicated the intercept and slope parameters of all of the students' level 1 models based on the invariant covariates (O'Connell & McCoach, 2008). The covariates were student characteristics: gender, social economic status (SES), English as second language program participating status (ESL), disability status (SWD), and race. Moreover, some researchers argue that the model needs to capture the complexities of student scores across years that might be attained in different schools by including school as one invariant covariate (McCaffrey et al., 2004). Although students usually stay in the same school in one academic year, schools implement different curriculum plans, which increase the complexities of growth modeling. Hence, school was added as an invariant covariate on the level 2 model to capture the school effects. All the invariant covariates information was based on the data collected in the 2011–2012 academic year. Then, the level 2 model is:

$$\pi_{0i} = \beta_{00} + \beta_{01}(\text{Gender}_i) + \beta_{02}(\text{SES}_i) + \beta_{03}(\text{ESL}_i) + \beta_{04}(\text{SWD}_i) + \beta_{05}(\text{Race}_i) + \beta_{06}(\text{School}_i) + r_{0i}, \quad (2)$$

$$\pi_{1i} = \beta_{10} + \beta_{11}(\text{Gender}_i) + \beta_{12}(\text{SES}_i) + \beta_{13}(\text{ESL}_i) + \beta_{14}(\text{SWD}_i) + \beta_{15}(\text{Race}_i) + \beta_{16}(\text{School}_i) + r_{1i}, \quad (3)$$

where π_{0i} is the intercept of the level 1 regression equation for student i; π_{1i} is the slope of the level 1 regression equation for student i; β_{00} and β_{10} are the intercept parameters of the level 2 regression equations representing the population's true initial status and growth rate, respectively, when all the predictors are zero; β_{01}, β_{02}, β_{03}, β_{04}, β_{05}, and β_{06} are the slopes of invariant covariates of the level 2 model representing the effect of these variables on the variation of students; β_{11}, β_{12}, β_{13}, β_{14} , β_{15}, and β_{16} are the slopes of invariant covariates of the level 2 model

representing the effect of these variables on the variation of students' growth rates; and r_{0i} and r_{1i} are the random error terms.

The level 2 model that allows the intercept and slope of the level 1 model to change among students is called the random intercept and random slope model (equations 2 and 3). This model is usually compared with the random intercept model (π_{1i} is fixed for all students, equation 2) that only allows the intercept of the level 1 model to change on the model fit index. The purpose of this comparison is to investigate whether different slopes among students' level 1 model make the model fit better. If the random intercept and random slope model fits the data better, π_{1i}, the slope of the level 1 linear equation, is the growth score for each student, which can be calculated using the values of invariant covariates and estimated parameters (β_{10}, β_{11}, β_{12}, β_{13}, β_{14}, β_{15}, and β_{16}) of equation 3.

Comparison Criteria

The TC and CY growth scores were compared on student growth score estimates and score precision. First, the authors computed both the correlation between the two growth scores and the ANCOVA test of the two growth scores, controlling the 2011–2012 grade 7 end-of-year summative test score and student characteristic variables. Then, an examination of the histogram of the two growth scores' difference revealed the extent of growth score differences. To compare the score precision, the authors used the estimated standard error of the two-level hierarchical linear model. The estimated standard error of the two-level hierarchical linear model showed the estimated precision of the two growth score estimates. The estimated precision is a key statistic of student growth score estimates as it provides information to determine the accuracy of a school's or teacher's performance evaluation using student growth scores (Goldschmidt, Choi, & Beaudoin, 2012).

Results

Table 2 displays the description statistics of the seven assessment scores. The mean scores of the four TC summative assessments and the four CY summative assessments all increased through time. However, the rate of increase of the four TC summative assessment scores was much higher than the rate of increase of the four CY summative assessment scores. The standard deviations of the three 2011–2012 interim assessment scores were very similar but were bigger than the standard deviation of the 2011–2012 end-of-year assessment score. The standard deviations of the four CY summative assessment scores decreased through the years.

With the HLM growth model, the covariates were student characteristics. As such, the frequencies of all student characteristics from the sample population of 1,439 students were also calculated. The results are collected in Table 3.

The parameters of the HLM growth models were estimated using the PROC MIXED procedure in SAS 9.3. Four models were fitted to the data in order to pick the better-fitted models. These four models were the random intercept model (equations 1 and 2) for both the TC assessment and the CY assessment and the random intercept and random slope model (equations 1, 2 and 3) for both the TC assessment and the CY assessment. The statistics used to judge model fit include the –2×log-likelihood, the Akaike information criterion (AIC), and the Bayesian information criterion (BIC). Table 4 includes the fit statistics for these four models.

To decide whether to use the random intercept model or the random intercept and random slope model for the growth scores, the two models were compared statistically. For the TC assessments, the log-likelihood ratio test between the random intercept model and the random intercept and random slope model was significant, $\Delta\chi^2(\Delta df = 8) = 148.07$, $p =< 0.01$. For the CY assessments, the log-likelihood ratio test between the random intercept model and the random

intercept and random slope model was also significant, $\Delta\chi^2(\Delta df = 8) = 175.43$, $p =< 0.01$. The AIC and BIC fit statistics decreased from the random intercept model to the random intercept and random slope model for both the TC assessments and the CY assessments. The significant log-likelihood ratio tests and decreasing fit statistics all indicated the random intercept and random slope model fit the data better for both the TC assessments and the CY assessments. Thus, the random intercept and random slope model was used to calculate the students' TC and CY growth scores.

The correlation between growth scores, the ANCOVA test, and the histogram all show that the differences between the students' TC and CY growth scores were large and that the students' TC growth scores were higher than their CY growth scores. The correlation between students' TC growth scores and CY growth scores was only 0.20. The overall ANCOVA model for comparing the two growth scores after controlling for the 2011–2012 grade 7 end-of-year summative test score and student characteristic variables was significant, $F(8, 2869) = 2297.95$, $p < 0.001$. The adjusted R-squared of this ANCOVA model was 0.87. The parameter estimation of this ANCOVA model indicated that the TC growth score is about 6.74 higher than the CY growth score on average, and this difference is significant, $t_{2869} = 133.88$, $p < 0.001$. Figure 1 shows the distribution of the two growth score differences.

Parameters used to calculate student growth scores were β_{10}, β_{11}, β_{12}, β_{13}, β_{14}, β_{15}, and β_{16}. These parameters are the slopes of time, time and gender interaction (time × gender), time and SES interaction (time × SES), time and ESL interaction (time × ESL), time and SWD interaction (time × SWD), time and race interaction (time × race), and time and school interaction (time × school), respectively. The estimated standard errors of these parameters from two growth models

are in Table 5. The standard errors of these parameters for the TC assessments were smaller than the standard errors of these parameters for the CY assessments.

Conclusion and Implications

The random intercept and random slope HLM model was used to calculate students' TC and CY growth scores. The differences between the students' TC growth scores and their CY growth scores were large, and the students' TC growth scores were higher than their CY growth scores. More negative growth scores were calculated using the CY summative assessment scores than using the TC summative assessment scores. Since the CY assessments were linked but not vertically scaled, the mean differences among the CY summative assessment scores were smaller than the mean differences among the TC summative assessment scores. This might be the reason that the CY growth scores were much lower than the TC growth scores. In future studies, vertically scaled CY assessment scores might lead to higher CY growth scores and smaller differences between TC and CY growth scores. For the model precision comparison, the lower parameter standard errors of the TC assessment growth model indicated higher precision of the TC growth score when compared with the CY growth score. Thus, this study provides evidence that the TC summative assessment is a better way to measure growth.

References

Goldschmidt, P., Choi, K., & Beaudoin, J. P. (2012). *Growth model comparison study: Practical implications of alternative models for evaluating school performance.*. Retrieved from the Council of Chief State School Officers website: http://www.ccsso.org/Documents /2012/Growth_Model_Comparison_Study_Evaluating_School_Performance_2012.pdf

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67-101. doi:10.3102/10769986029001067

Mislevy, R. J. & Zwick, R. (2012). Scaling, linking, and reporting in a periodic assessment system. *Journal of Educational Measurement, 49*(2), 148-166. doi:10.1111/j.1745-3984.2012.00166.x

O'Connell, A. A., & McCoach, D. B. (Eds.). (2008). *Multilevel modeling of educational data*. Greenwich, CT: Information Age Publishing.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

U. S. Department of Education. (2010). Overview information; Race to the Top Fund Assessment Program; Notice inviting applications for new awards for fiscal year (FY) 2010. 75 Fed. Reg. 18171–18185 (Apr. 9, 2010).

Table 1

Sample Count for the Equipercentile Linking Method

Linked tests	Students
2011 and 2010 EYA	29,533
2011 and 2009 EYA	28,230
2011 and 2008 EYA	28,945

Note. EYA = End of school year assessment.

Table 2

Mean and Standard Deviation of Seven Assessment Scores

Measure	Interim 1	Interim 2	Interim 3	EY 2009	EY 2010	EY 2011	EY 2012
Mean	56.30	61.72	69.05	76.25	76.60	76.39	77.19
SD	16.39	15.88	15.45	14.27	13.84	13.05	12.62

Note. EY = End of school year.

Table 3

Frequency of Student Characteristics in the Growth Score Sample

Characteristic	*n*
Gender	
Girls	709
Boys	730
Socioeconomic Status	
Not Eligible	936
Reduced Lunch Eligible	146
Free Lunch Eligible	357
Disability	
No	1412
Yes	27
ESL Program	
No	1401
Yes	38
Ethnicity	
Native American	10
Asian	21
African American	82
Hispanic	144
Caucasian	1103
Multi-Ethnic	78
Pacific Islander	1

Note. The sample in question is the growth score sample. There were 1,439 students in 39 different schools.

Table 4

Fit Statistics of Four Models

	TC Assessment		CY Assessment	
Fit Statistic	RI	RIRS	RI	RIRS
–2×log likelihood	43549.78	43401.71	42912.71	42737.28
AIC	43569.81	43437.83	42932.74	42773.4
BIC	43622.49	43532.6	42985.42	42868.17

Note. TC = through-course; CY = cross-year; RI = random intercept model; RIRS = random intercept and random slope model.

Table 5

Standard Errors of Random Intercept and Random Slope Growth Model Parameters

	Standard Error	
RIRS Parameters	TY Assessment	CY Assessment
Time	0.62	0.66
Gender × Time	0.20	0.21
SES × Time	0.12	0.13
ELS × Time	0.64	0.68
SWD × Time	0.73	0.78
Race × Time	0.14	0.15
School × Time	0.00	0.00

Note. TC = through-course; CY = cross-year; RIRS = random intercept and random slope model; SES = socioeconomic status; ELS = English as a second language program participation; SWD = disability status.

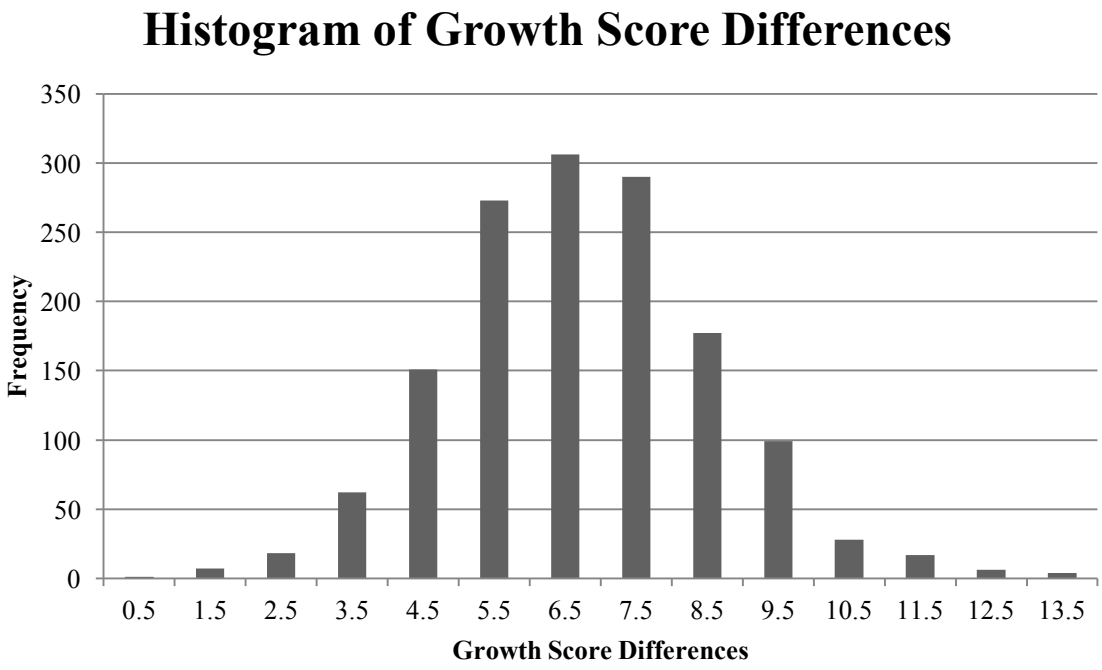

Figure 1. Histogram of growth score differences. Differences are calculated by subtracting the CY growth score from the TC growth score.