So, what d'ya expect? Pursuing Reasonable Individual Student Growth Targets to Improve Accountability Systems

Carl Hauser
Northwest Evaluation Association

A paper for the symposium: *Issues in the use of Longitudinal Student Achievement Data in Monitoring School Effectiveness and School Improvement*

Presented at the 2003 Annual Meeting of the American Educational Research Association, Chicago, IL

So, what d'ya expect? Pursuing Reasonable Individual Student Growth Targets to Improve Accountability Systems

Carl Hauser
Northwest Evaluation Association
April 2003

One unfortunate characteristic of most highly visible educational accountability systems is their close tie to a single or very few consequential levels of academic achievement. For example, the Adequate Yearly Progress provision of the *No Child Left Behind Act* of 2001 focuses exclusively on a "proficiency" level of achievement. Since attainment of "proficiency" is the sole level for which credit is granted in this system, the concern is that students making good progress but not enough to be considered "proficient" receive no credit toward an index of being "accountable". Over time, students who are considered too far from the "proficiency" level to be able to attain it by assessment time, run the risk of losing instructional attention in favor of more "proficient probable" students. Similarly, students who are obviously beyond the key "proficiency" level can also lose instructional attention. Any positive change in their status will not affect the accountability index.

Various authors have addressed this dilemma. For example, Linn, Baker, and Betebenner (2002) proposed a system that assigns fractional credit to performance categories other than "proficient". Flicek and Lowham (2001) proposed using individual student growth referenced to longitudinal growth norms as a method of incorporating and giving credit for progress made, even though the end performance status might fall short of a performance criterion. Kingsbury (2000) proposed a "hybrid success model" for setting individual student growth expectations for students based on their proximity to an achievement target, thus allowing both status and growth to demonstrate accountability. What distinguishes the Flicek and Lowham and the Kingsbury proposals is the central role each assigns to individual student growth within an accountability scheme. Both consider individual growth as an integral part to accountability, not merely as an optional supplement (or worse, an interesting side note) to performance status.

This paper is predicated on a rather simple argument: in order for academic growth to serve in a fundamental role in an accountability system, the amount of growth a student would *reasonably* be expected to attain over some set time interval (i.e., a growth expectation, standard, or target) must be able to be declared in advance. These declarations, or others based on them, will typically be translated into a form of value within an accountability scheme. This value, in turn, will be at least part of the evidence for judging the extent to which the school (or district or state) is being successful or 'accountable'. For example, a district expectation might be that all 4th grade students grow by X amount. While this expectation is certainly convenient, its reasonableness is open to question. Is it reasonable to assume that all students in a single grade would grow at the same rate? For a high achieving student, requiring average grade level growth will likely be more demanding than requiring average grade level growth from a lower achieving student. A lower achieving student might even be thought of as "under-challenged". Neither student would be treated equitably.

A 'reasonable' growth target can be thought of as the proximity between the observed growth and the expected growth; the closer the observed growth is to expected growth, the more reasonable the growth target. This position implies that observed growth that is substantially greater than the target is no more or less reasonable than observed growth that is substantially less than the target. With a focus on *individual* student growth, it should be possible to create a method of defining reasonable, equitable growth targets for each student using characteristics of the individual student's past performance. There is already strong evidence, for example, that the rate of growth is often associated with initial student achievement status (e.g., NWEA, 2002; Seltzer, Choi, & Thum, 2002a, 2002b).

The purpose of this study was to evaluate several feasible models for determining single-year academic growth targets for individual students. These models are detailed in the next section. Single-year growth targets were considered as the most likely points from which declarations of the value of observed growth would be defined for use in an accountability system (e.g., "value added" systems). This study was undertaken as an initial, empirical exploration of some of the territory involved in this area. The study is certainly not definitive, though it holds implications for questions such as: 'How much academic growth can we reasonably expect a student to make over the course of a year?'; 'Is it reasonable to ask all students in the same grade to grow at the same rate?'; 'Can the observed growth of large numbers of students who were in the same grade level and in the same achievement range, help to define reasonable growth?'. The study does not address *how* growth data, per se, should be used in an accountability system, only on how an equitable baseline of growth could be established.

Methods

Data sources.

Data for the study came from three cohorts of student test records. Two of the cohorts (A and B) were from a single moderate sized school district in Wyoming. The district has 28 elementary and four middle schools and a total student population of slightly over 12,000. For these cohorts there were four waves each of spring achievement data in reading and mathematics (spring 1999 through spring 2002). The third cohort (C) came from the Northwest Evaluation Association *2002 RIT Scale Norms Study*. The test records making up this cohort are from students in nine districts in six states. In this set there were 10 waves of fall and spring achievement data in reading and mathematics (fall 1996 through spring 2001). For Cohorts A and B, the last wave contained the scores to be predicted. In Cohort C, the last wave also contained the scores to be predicted. But in Cohort C, the ninth wave (fall 2000) was not considered as observed. In all cohort datasets, only those student records containing complete test data for a subject area were included in the analyses for that area. Thus, for example, a particular student's complete reading test data would be included even though their mathematics test data were incomplete (and not included). These cohort characteristics are summarized in Table 1.

Table 1. Characteristics of the cohort data sets

Cohort	Districts represented	Total waves	"Observed" waves used for prediction term,year [a]	Predicted term,year / grade	Reading	Math
A	1	4	S99, S00, S01	S02 / 5	655	659
B	1	4	S99, S00, S01	S02 / 6	738	742
C	9	10	F96, F97, F98, F99, S97, S98, S99, S00	S01 / 8	3876	4132

[a] F = fall; S = spring

Table 2 presents achievement data for the three cohorts. Achievement levels between the cohorts were comparable in common grades for the spring terms. Variance in common grades in Reading tended to be slightly higher in Cohorts A and B than for Cohort C. The reverse was true in Mathematics. In Mathematics for Cohort C, a trend of increasing variance from the first wave to the last was observed.

Table 2. Descriptive statistics of cohort performance in reading and mathematics by season and year.

Season-Year	Grd	Reading Med.	Reading Mean	Reading SD	Reading Min.	Reading Max.	Reading N	Matematics Med.	Matematics Mean	Matematics SD	Matematics Min.	Matematics Max.	Matematics N
Cohort A													
S-99	2	192	188.7	14.72	148	226	655	190	189.0	12.54	144	226	659
S-00	3	202	199.0	13.81	152	234	655	203	201.6	12.42	148	239	659
S-01	4	210	207.3	14.03	144	239	655	213	212.2	11.07	159	250	659
S-02*	5	215	213.9	12.78	164	252	655	222	221.9	12.15	174	264	659
Cohort B													
S-99	3	201	198.5	13.83	148	232	738	202	200.5	11.57	161	229	742
S-00	4	209	206.4	12.77	152	237	738	211	210.9	11.58	168	255	742
S-01	5	215	213.7	12.95	154	247	738	220	220.2	12.45	177	254	742
S-02*	6	220	219.1	12.40	155	247	738	228	226.8	12.73	180	262	742
Cohort C													
F-96	4	204	201.7	13.49	143	233	3876	201	200.4	11.30	149	247	4132
S-97	4	210	208.3	13.20	143	243	3876	210	209.4	12.13	154	255	4132
F-97	5	211	209.1	12.96	147	241	3876	210	209.5	12.42	155	252	4132
S-98	5	216	214.6	12.67	154	251	3876	218	218.0	12.97	150	263	4132
F-98	6	217	215.5	12.12	155	250	3876	218	217.2	13.10	172	261	4132
S-99	6	222	220.1	11.97	156	258	3876	225	225.5	14.88	172	278	4132
F-99	7	222	220.1	11.68	156	256	3876	226	226.2	15.25	160	282	4132
S-00	7	225	223.5	12.37	166	261	3876	234	233.9	16.35	171	293	4132
F-00	8	225	224.3	11.95	160	268	3876	235	234.5	16.56	164	290	4132
S-01*	8	230	229.1	11.60	165	269	3876	242	242.1	17.10	184	294	4132

* Designates the season-year for which scores were predicted.

Tests characteristics. All tests used in this study were created from the NWEA item banks in Reading and Mathematics. These banks are comprised of several thousand test items that have been calibrated for difficulty using the one-parameter Item Response Theory (IRT) model (Rasch model). Item difficulty and student ability are both expressed in Rasch Units (RITs) on the same scale. A RIT is simply the linear transformation of the logit theta metric that sets the unit at .10 logits and centers the scale at 200 (i.e., RIT = θ*10 + 200). Thus, a RIT of 210 is equivalent to logit = 1. There is one scale for Reading and one scale for Mathematics. Paper and pencil Achievement Level Tests in Reading can measure dependably from about RIT 149, ±3.6 (percentile 2 in fall grade 2) to about RIT 252, ±5.1 (percentile 98 in spring grade 10). In Mathematics, paper and pencil tests measure accurately from about RIT 156, ±3.8 (percentile 2, fall grade 2) to about RIT 276, ±5.5 (> percentile 98 in spring grade 10). Well-targeted level tests typically have measurement error in the 2.8 – 3.3 range. Computerized-adaptive versions extend slightly the measurement ranges with these levels of associated measurement error. A complete description of the technical characteristics of NWEA tests can be found in the *NWEA Technical Manual for Achievement Level Tests and Measures of Academic Progress* (2003).

NWEA RIT Scale Norms. Several of the models used to determine individual student growth targets used data reported in the NWEA 2002 norms study. This study includes the test records of approximately 1.05 million students representing 321 school districts in 24 states. The districts ranged from very urban to very rural. They ranged in size from under 200 to over 60,000 students.

The norms study provided several specific data elements. Grade level means and standard deviations of student status and growth in the grades of interest were used. For status level data, these were based on roughly 71,000 to 89,000 students per grade level. Grade level growth means were based on intact groups of students; that is, student growth was based on the same students having both scores used to calculate a change (growth) score. Spring-to-spring grade level growth means were based on roughly 44,000 to 54,000 students per grade level. Growth means were also retrieved that were disaggregated by the starting status level of students. These means were calculated for all students whose achievement status at the beginning of the comparison period fell into each 10 point RIT block. RIT blocks were set at 140-149, 150-159, 160-169, ... ≈ 260-269. The numbers of students used to compute the means in these RIT block cells ranged from 258 to over 14,000. Average N's for all RIT block cells were 4427 for Reading and 4495 for Mathematics. Spring-to-spring growth distributions are summarized in Tables 3a, 3b for Reading and in Tales 4a, and 4b for Mathematics.

Table 3a. Means and standard deviations of spring-to-spring achievement growth in Reading by grade level and initial RIT block

	140-149		150-159		160-169		170-179		180-189		190-199		200-209		210-219		220-229		230-239		240-249	
Grd	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
2-3	***17.3***	***11.31***	16.0	11.08	16.1	9.66	13.9	8.20	12.5	7.50	11.0	6.40	9.2	6.06	7.2	5.74						
3-4			13.8	9.80	14.0	9.59	11.7	8.52	9.9	7.42	8.5	6.65	7.0	6.03	5.5	5.91	3.5	5.95	-2.6	7.47		
4-5			12.2	10.19	13.3	8.73	11.6	8.37	9.9	7.63	8.4	6.88	6.9	6.10	6.0	5.74	4.5	5.56	3.2	6.14		
5-6					12.3	9.02	10.3	8.56	9.2	7.99	7.6	7.19	6.2	6.22	5.4	5.76	3.9	5.37	2.6	5.67	-0.6	6.34
6-7					***10.3***	***10.01***	9.6	8.57	8.0	8.22	6.8	7.39	5.6	6.50	4.8	5.91	3.7	5.47	2.8	5.65	1.7	5.93
7-8							10.1	8.26	8.9	8.36	7.5	7.90	6.4	6.65	5.2	6.04	3.9	5.54	2.8	5.62	1.6	6.21
8-9											6.7	7.93	5.1	7.39	4.1	6.48	2.9	5.70	1.7	5.51	0.4	6.22
9-10															***3.7***	***6.94***	3.3	6.96	3.0	6.60		

Note: ***Bold italized*** entries indicate 250-299 students

Table 3b. Number of students used in the calculation of spring-to-spring growth estimates in Reading by grade level and initial RIT block

Grd	140-149	150-159	160-169	170-179	180-189	190-199	200-209	210-219	220-229	230-239	240-249
2-3	285	1186	1573	2830	3963	4201	2973	890			
3-4		988	1674	3258	6813	10598	13236	8427	2626	562	
4-5		468	886	2006	4525	8437	13446	13247	6983	1448	
5-6			491	1053	2567	5626	10346	13979	10825	3900	551
6-7			265	607	1699	3839	8167	13261	14273	7281	1290
7-8				356	935	2208	5172	9379	12951	8629	1954
8-9						370	1092	2031	3176	2348	620
9-10								282	558	428	

Table 4a. Means and standard deviations of spring-to-spring achievement growth in Mathematics by grade level and initial RIT block

	150-159		160-169		170-179		180-189		190-199		200-209		210-219		220-229		230-239		240-249		250-259		260-269	
Grd	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
2-3	17.9	10.09	14.9	9.58	14.1	8.99	13.2	7.56	12.6	6.70	11.0	6.29	9.7	6.48										
3-4	***17.2***	***9.89***	14.0	9.74	12.5	8.42	10.8	7.42	9.6	6.63	8.9	6.42	8.5	6.57	8.4	6.70	7.6	7.19						
4-5			13.9	9.58	12.0	8.30	10.0	7.50	9.4	6.74	9.1	6.45	9.3	6.28	9.5	6.22	8.8	6.31	8.2	6.36				
5-6					8.7	8.44	7.6	7.79	6.1	7.40	6.1	6.97	6.5	6.70	6.9	6.50	7.5	6.59	6.7	6.67	4.3	7.03		
6-7					9.7	8.52	7.7	7.57	6.0	6.91	6.2	6.77	7.0	6.42	7.3	6.32	8.0	6.47	7.6	6.38	6.3	7.11	4.0	8.14
7-8							8.1	7.71	7.0	7.25	8.0	7.80	8.4	7.25	8.8	7.05	9.1	6.75	8.5	6.93	7.3	7.41	4.1	8.06
8-9											10.9	12.53	11.9	10.80	11.2	9.31	10.4	8.08	7.2	7.32	3.6	7.68	0.3	8.48
9-10																	7.7	9.90	3.8	7.84	1.1	7.66	-1.8	8.61

Note: ***Bold italized*** entries indicate 250-299 students

Table 4b. Number of students used in the calculation of spring-to-spring growth estimates in Mathematics by grade level and initial RIT block

Grd	**150-159**	**160-169**	**170-179**	**180-189**	**190-199**	**200-209**	**210-219**	**220-229**	**230-239**	**240-249**	**250-259**	**260-269**
2-3	358	1136	3085	5620	5331	2726	665					
3-4	258	723	2577	6267	12313	15148	8997	2530	450			
4-5		317	1157	3212	7831	13745	13763	7228	2188	448		
5-6			518	1593	4277	9291	13318	12204	6743	2569	498	
6-7			359	1139	3094	6791	10724	12488	9967	5307	1862	361
7-8				659	1954	4036	6382	8842	9109	6595	3498	1049
8-9						496	1063	1934	2770	3093	1852	749
9-10									446	1139	1618	819

Models for determining individual student growth targets.

All the models investigated yielded a prediction of each student's final term RIT score in the subject area being considered. For Cohorts A and B, this was spring 2002, for Cohort C it was spring 2001. Individual student prediction residuals (observed score – predicted score) were used as the basis for comparing the models. The models differed in the way the available data (prior to the final term) were treated and combined with a growth estimate to arrive at a prediction. Models based on mean *z*-score status were the only models not to include an explicit estimate of growth. Some models used growth norm references from the 2002 NWEA norming study. One model used no prior achievement data but only the mean observed growth of same grade-level students from the norms study. A second model used only the observed RIT score from the spring prior to the final (predicted) term and the mean observed growth of students who achieved a similar RIT score at the same grade level from the norms study. All other models used all prior RIT scores from a student's record to arrive at a growth estimate for the student. Some used the scores directly while others relied on modeling these scores to "true" score estimates using linear modeling (LM). Stated more formally, the models are as follows:

Mean grade level growth (MGLG):

$$\hat{Y}_{g+1} = RIT_{gi} + \mu_{g,}$$

Where RIT_{gi} is the observed RIT score for student *i* in grade *g*, the final observed grade; μ_g is the mean growth of students in the norms study going from grade *g* to *g+1*.

Mean RIT block growth (MRBG):

$$\hat{Y}_{g+1} = RIT_{gi} + \mu_{RB.g}$$

Where RIT_{gi} is the observed RIT score for student *i* in grade *g*, the final observed grade; μ_{RB} is the mean RIT block growth of students in the norms study going from *g* to *g+1* whose achievement in the final observed grade, *g* was in RIT block, *RB*.

Linear Model (LM) least squares slope estimate (LMlsSlp):

$$\hat{Y}_{g+1} = RIT_{gi} + \pi_{1i.LS}$$

Where RIT_{gi} is the observed RIT score for student *i* in grade *g*, the final observed grade; $\pi_{1i.LS}$ is the LM least squares estimate of growth rate for student *i* over the entire data collection period.

In the LMLSslp model and all the models below that include a linear model (LM) component, the linear model component developed was equivalent to the level 1 model of a hierarchical linear model

(HLM). The level 1 model was structured as it might be posed in a study of academic achievement growth in a school system; that is, without predictor variables and using grade level as the time variable. In contrast to a growth study, however, the time variable, grade, was 're-centered' on the last observed grade so that it took on a value of 0 while prior grades took on negative values. For example in Cohort A, grades 2, 3, and 4 became grades -2, -1, and 0 respectively used to predict grade 5 which took on the value of +1. When fall scores were included in the analyses (Cohort C), the decimals .1 and .8 were used to distinguish between fall and spring, respectively. Centering on the final 'observed' grade (7.8) resulted in the grades 4.1, 4.8, 5.1, 5.8, 6.1, 6.8, 7.1, and 7.8 being converted to -3.7, -3.0, -2.7, -2.0, -1.7, -1.0, -.7, and 0, respectively. All models that included linear components were estimated using HLM5 (Raudenbush, Bryk, Cheong & Congdon, 2001).

It should also be noted that the Cohort C data were analyzed using a linear and a non-linear (quadratic) model in order to evaluate best model fit. These analyses supported the use of a linear model over a non-linear model for both Reading and Mathematics using spring only and fall and spring data.

Linear Model (LM) empirical Bayes slope estimate (LMeBSlp):

$$\hat{Y}_{g+1} = RIT_{gi} + \pi_{1i.EB}$$

Where RIT_{gi} is the observed RIT score for student *i* in grade *g*, the final observed grade; $\pi_{1i.EB}$ is the LM empirical Bayes estimate of growth rate for student *i* over the data collection period.

Linear Model (LM) least squares status estimate with RIT block growth (LMlsSt+MRBG):

$$\hat{Y}_{g+1} = \pi_{0i.LSg} + \mu_{RB.g}$$

Where $\pi_{0i.LSg}$ is the LM least squares estimate of the status for student *i*, in grade *g*, the final observed grade; $\mu_{RB.g}$ is the mean growth of students in the norms study going from grade *g* to *g+1* whose achievement in grade, *g*, was in RIT block, *RB*.

Linear Model (LM) empirical Bayes status estimate with RIT block growth (LMeBSt+MRBG):

$$\hat{Y}_{g+1} = \pi_{0i.EB} + \mu_{RB.g}$$

Where $\pi_{0i.EB}$ is the LM empirical Bayes estimate of the status for student *i* in grade *g*, the final observed grade, *g*; $\mu_{RB.g}$ is the mean growth of students in the norms study going from grade *g* to *g+1* whose achievement in grade, *g*, was in RIT block, *RB*.

Full Linear Model (LM) least squares status and growth rate estimates (FLMls):

$$\hat{Y}_{g+1} = \pi_{0i.LS} + \pi_{1i.LS}g_{ti} + e_{ti}$$

Where $\pi_{0i.LS}$ is the LM least squares estimate of the status for student *i* when the grade metric, g_{ti} = 0; $\pi_{1i.LS}$ is the LM least squares estimates of the growth rate for student *i* over the data collection period; e_{ti} is error. The final observed grade, *g*, was set to g = 0, and all prior grades were reset according to g-1 = -1, g-2 = -2, and so on.

Full Linear Model (LM) empirical Bayes status and growth rate estimates (FLMeB):

$$\hat{Y}_{g+1} = \pi_{0i.EB} + \pi_{1i.EB} g_{ti} + e_{ti}$$

Where $\pi_{0i.EB}$ is the LM empirical Bayes estimate of the status for student *i* when the grade metric, $g_{ti} = 0$; $\pi_{1i.EB}$ is the LM least squares estimates of the growth rate for student *i* over the data collection period; e_{ti} is error. The final observed grade, *g*, was set to g = 0, and all prior grades were reset according to g-1 = -1, g-2 = -2, and so on.

Mean of norms-based z scores (MnbZ):

$$\hat{Y}_{g+1} = Z_{g+1} = \left[\frac{Z_{g-n} \ldots + Z_{g-3} + Z_{g-2} + Z_{g-1} + Z_g}{n}\right] * \sigma_{g+1} + \mu_{g+1}$$

Where z for the predicted grade, *g+1*, is the mean of norm-based z's from all prior tests using the respective means and standard deviations, as found in the norms study, from the earliest grade, *g-n*, to the final observed grade, *g*, and σ_{g+1} and μ_{g+1} are the standard deviation and the mean, respectively of the grade *g+1* from the norms study.

Mean of norms-based z scores with last observed score double weighted (MnbZ*):

$$\hat{Y}_{g+1} = Z_{g+1} = \left[\frac{Z_{g-n} \ldots + Z_{g-3} + Z_{g-2} + Z_{g-1} + 2Z_g}{n+1}\right] * \sigma_{g+1} + \mu_{g+1}$$

Where z for the predicted grade, *g+1*, is the mean of norm-based z's from all prior tests using the respective means and standard deviations, as found in the norms study, from the earliest grade, *g-n*, to the final observed grade, *g* which is double-weighted, and σ_{g+1} and μ_{g+1} are the standard deviation and the mean, respectively of the grade *g+1* from the norms study.

Mean of locally based z scores (MlbZ):

$$\hat{Y}_{g+1} = Z_{g+1} = \left[\frac{Z_{g-n} \ldots + Z_{g-3} + Z_{g-2} + Z_{g-1} + Z_g}{n}\right] * sd_{g+1} + \overline{X}_{g+1}$$

Where z for the predicted grade, *g+1*, is the mean of locally based z's from all prior tests using the means and standard deviations calculated from scores in the earliest grade, *g-n*, to the final observed grade, *g*, and sd_{g+1} and $\overline{X}_{g+1}$ are the local historical standard deviation and the mean, respectively of grade *g+1*.

Mean of locally based z scores with last observed score double weighted (MlbZ*):

$$\hat{Y}_{g+1} = Z_{g+1} = \left[\frac{Z_{g-n} \;...+\; Z_{g-3} + Z_{g-2} + Z_{g-1} + 2Z_g}{n+1}\right] * sd_{g+1} + \overline{X}_{g+1}$$

Where z for the predicted grade, *g+1*, is the mean of locally based z's from all prior tests using the means and standard deviations calculated from scores in the earliest grade, *g-n*, to the final observed grade, *g*, which is double-weighted, and sd_{g+1} and $\overline{X}_{g+1}$ are the local historical standard deviation and the mean, respectively of grade *g+1*.

All models except for the last two were applied to the cases in all three cohorts. All data from each set were used, with the last score used in prediction (referred to above as the last observed grade, *g*) being the score from the spring one year prior to the spring score being predicted (i.e., grade *g+1*). This means that for Cohort C, where the RIT being predicted was for grade 8, the *fall* grade 8 RIT was not used in any of the prediction models. The last two models, locally-based *z* scores, could only be applied to the Cohort A data for two reasons: a) data for Cohort C were collected across districts, thus common local means and standard deviations were not available, and b) no historical local data were available to supply the means and standard deviations for the predicted grade for Cohort B, grade 6.

Analysis.

Residuals at the individual student level ($Y_{g+1} - \hat{Y}_{g+1}$) yielded from each of the models were the focus of analysis. For each set of predictions from each cohort, several statistics were computed to help describe the resulting distribution of residuals. These included the mean residual, the root mean square error, and the percent of the cases for each model that yielded the minimum residual across all models. To assess how well each model's uniformity in prediction across the measurement range, Pearson product-moment correlations were calculated between the residuals and the last observed RIT score. Positive correlations indicate that higher scores will tend to be under-predicted and lower scores will tend to be over-predicted. Negative correlations indicate the opposite tendencies. The extent of these deviations depends on the magnitude of the correlation. In addition, the percent of cases for each model that yielded a predicted score within a reasonable standard error band of the observed score was calculated. 'Reasonable', here, was considered to be ±3.3 for Reading and ±3.2 for Mathematics. These values were based on examinations of the error levels observed for well targeted tests – raw score 45-65 percent correct. Comparisons between methods were also maintained at the descriptive level. More specifically, plots of residuals by the final (observed) score were developed to form a more complete understanding of the nature of prediction results of the various models.

Results

Cohorts A and B.

Table 5 contains the results of the five basic descriptive statistics for Cohorts A and B for both Reading and Mathematics. The asterisks in Table 5 designate the most favorable value for the particular descriptive statistic across all models. Similarly, the superscript italic 2's designate the next most favorable value for the statistic. For example, in the Cohort A – Reading results, the MlbZ* model was found to have the most favorable mean residual (minimum absolute) value (.18), while the MRBG model was the next most favorable (-.18).

When examining Table 5 within each content area, several commonalities appear. Initially we see that the linear models that included the slope parameter (LMlsSlp, LMeBSlp, FLMls, and FLMeB) in the prediction, tended to result in an over-prediction bias indicated by large negative values of the mean residual. This unfavorable outcome was evident in each of the other indicators. Models involving mean RIT block growth (MRBG, LMlsSt+MRBG, and LMeBSt+MRBG) resulted in somewhat more favorable results across indicators for Cohort A in both Reading and Mathematics. In fact the linear model using empirical Bayes estimates of status with RIT block growth as estimates of rate (LMeBSt+MRBG) produced the most favorable results in Reading. In Mathematics, however, the model using local-based z scores (MlbZ and MlbZ*) produced the most favorable set of results even though results generally under-predicted performance.

For Cohort B, linear models that included the estimation of grade status from a linear model in combination with RIT block growth means as estimates of rate of growth, LMlsSt+MRBG (for Reading) and LMeBSt+MRBG (for Mathematics) yielded the most favorable set of indicators. In terms of percentage of predictions within the 1SEM bands established, the norm-based *z*-score models (MnbZ and MnbZ*) were both favorable for Reading. In Mathematics, the simple models using only mean grade level growth (MGLG) and RIT block growth (MRBG) were also quite favorable. In both cases, however, the correlations between the residual and the last observed RIT score were too high for these models to be considered across the measurement range.

Table 5. Achievement Status Residuals by Method - Cohorts A and B

Grades 2-4 Spring Data Predicting Grade 5 Spring Status (Cohort A)

		Reading					Mathematics				
Model	Description	Mean resid.	RMSE	$r_{\hat{Y}res.g}$	% with min. resid.	% ±1 SE	Mean resid.	RMSE	$r_{\hat{Y}res.g}$	% with min. resid.	% ±1 SE
MGLG	Obs end Grd Status + Mean Gth	-0.45	7.527	-.419	5.6	38.6	0.48	6.640	-.130	4.6	37.0
MRBG	Obs end Grd Status + RIT Blk Mean Gth	-0.18 [2]	7.018	-.189	8.1	41.1	0.37 [2]	6.582	-.109	4.2	37.3
LMlsSlp	Obs end Grd Status + LM OLS est. Slope	-2.74	9.879	-.419	7.6	28.5	-1.95	9.041	-.192	7.4	27.6
LMeBSlp	Obs end Grd Status + LM EB est. Slope	-2.74	8.136	-.406	6.7	34.5	-1.95	6.807	-.046	9.6	37.9
LMlsSt+MRBG	LM OLS est end Grd Status + RIT Blk Mean Gth	-0.51	6.885	-.144	7.0	41.8	0.04 *	6.386	-.092 [2]	6.5	38.2
LMeBSt+MRBG	LM EB est end Grd Status + RIT Blk Mean Gth	-0.51	6.491 *	.066 *	11.6 [2]	42.6	0.04 *	6.256	.288	8.2	40.7
FLMls	Full LM OLS est end Grd Status & Slope	-3.07	9.776	-.389	6.7	27.8	-2.28	9.042	-.178	9.6	26.7
FLMeB	Full LM EB est end Grd Status & Slope	-3.07	7.965	-.231	11.6	37.3	-2.28	6.321	.345	17.3 *	33.7
MnbZ	Mean obs norm-based means, sd to predict z	1.32	6.864	-.216	18.0 *	42.9	3.07	6.695	-.236	16.5 [2]	35.4
MnbZ*	Mean obs norm-based means, sd (last double weighted) to predict z	1.14	6.740	-.277	5.5	42.6	2.68	6.301	-.264	7.2	37.6
MlbZ	Mean obs local-based means, sd to predict z	0.19	6.669	-.108 [2]	6.3	45.2 [2]	1.71	6.148 [2]	-.030 *	4.2	40.8 [2]
MlbZ*	Mean obs local-based means, sd (last double weighted) to predict z	0.18 *	6.538 [2]	-.162	5.2	45.8 *	1.70	5.992 *	-.110	4.2	41.3 *

Grades 3-5 Spring Data Predicting Grade 6 Spring Status (Cohort B)

		Reading					Mathematics				
MGLG	Obs end Grd Status + Mean Gth	0.08 [2]	6.266	-.327	8.4	41.2	0.18 [2]	6.014	-.186	11.5	42.0 [2]
MRBG	Obs end Grd Status + RIT Blk Mean Gth	0.08 [2]	5.966	-.076	5.4	41.7	-0.15 *	6.078	-.224	4.6	41.2
LMlsSlp	Obs end Grd Status + LM OLS est. Slope	-2.21	7.981	-.341	9.6	34.0	-3.28	7.849	-.339	6.2	29.0
LMeBSlp	Obs end Grd Status + LM EB est. Slope	-2.21	6.311	-.290	9.2	39.3	-3.28	6.440	-.297	8.0	33.0
LMlsSt+MRBG	LM OLS est end Grd Status + RIT Blk Mean Gth	0.01 *	5.953 [2]	-.014 *	7.7	44.9 *	-0.33	5.919	-.181	9.3	41.5
LMeBSt+MRBG	LM EB est end Grd Status + RIT Blk Mean Gth	0.01 *	6.171	.238	12.5 [2]	43.1	-0.33	5.705 [2]	.035 *	11.5	44.2 *
FLMls	Full LM OLS est end Grd Status & Slope	-2.28	8.086	-.290	8.1	34.0	-3.46	7.798	-.304	7.1	27.9
FLMeB	Full LM EB est end Grd Status & Slope	-2.28	5.850 *	.016 [2]	16.7 *	43.9	-3.46	5.651 *	-.062 [2]	15.4 [2]	37.2
MnbZ	Mean obs norm-based means, sd to predict z	1.70	6.186	-.103	16.7 *	41.7	2.20	6.462	-.200	18.2 *	35.8
MnbZ*	Mean obs norm-based means, sd (last double weighted) to predict z	1.41	5.956	-.155	5.7	44.2 [2]	1.82	6.113	-.254	8.4	38.9

Cohort C.

Cohort C results are contained in Table 6. The upper part of the table presents residuals based only on spring data while the lower part presents residuals based on both fall and spring data. For both Reading and Mathematics, the additional fall data had only minimal effect on bringing the mean residual closer to zero. For Reading, the inclusion of fall data into the two full linear models (FLMls and FLMeB), the actually introduced more bias into the predictions. However, the large over-prediction levels associated with linear models involving a slope parameter that were noted in the Cohorts A and B data were not as pronounced in the for Reading and were virtually absent for Mathematics.

Variance (RMSE) in the residuals of the models using linear estimates (LMlsSlp, LmeBSlp, LMlsSt+MRBG, LMeBSt+MRBG, FLMls, & FLMeB) was, in general, more favorable when both fall and spring data were used. This was the case for both Reading and Mathematics. Predictions in Reading using fall and spring and spring only data had the least variance when the full empirical Bayes linear model (FLMeB) and the linear model using empirical Bayes estimates of end grade status and RIT block mean growth for the rate estimate (LmeBSt+MRBG). In Mathematics, the linear model estimating end grade status using ordinary least squares and RIT Block mean for the rate estimate (LMlsSt+MRBG) and the simple observed end grade status plus mean RIT block growth (MRBG) resulted in the lowest levels of residual variance.

The linear models using empirical Bayes estimates of end grade status resulted in the most accurate (i.e., the highest percentage of cases within ±1 SEM) predictions in Reading when fall and spring data were used. For the spring only data, the full linear model using empirical Bayes estimates (FLMeB) resulted in the most desirable statistics overall. For the fall and spring data, the model using empirical Bayes and the model using ordinary least squares estimates of end grade status plus mean RIT block growth can be seen as the most. It lead to the most accurate predictions overall.

Prediction accuracy in Cohort C mathematics was highest for the simple observed end grade status plus mean RIT block growth model (MRGB) was found to be the most effective overall, even though some of its indicator statistics were not optimal. This was the case for both the spring only and the fall and spring datasets. However, the two linear models that used RIT block growth as the estimate of rate yielded accuracy percentages that approached that of the MRBG model in the fall and spring dataset. The norms-based *z-score* models yielded the least accurate predictions by far, particularly for the fall and spring dataset.

Table 6. Achievement Status Residuals by Method - Cohort C

Grades 4-7 Spring ONLY Data Predicting Grade 8 Spring Status

		Reading					Mathematics				
Model	Description	Mean resid.	RMSE	$r_{\hat{Y}res.g}$	% with min. resid.	% ±1 SE	Mean resid.	RMSE	$r_{\hat{Y}res.g}$	% with min. resid.	% ±1 SE
MGLG	Obs end Grd Status + Mean Gth	1.26	6.280	-.372	4.2	36.9	0.06 [2]	6.772	-.095	7.0	42.2 *
MRBG	Obs end Grd Status + RIT Blk Mean Gth	1.54	5.913	-.154	9.2	43.0	-0.10	6.674 *	-.013 *	11.3	39.9 [2]
LMlsSlp	Obs end Grd Status + LM OLS est. Slope	0.44 [2]	7.256	-.370	8.2	35.0	0.06 [2]	7.844	-.358	10.5	35.1
LMeBSlp	Obs end Grd Status + LM EB est. Slope	0.44 [2]	6.454	-.344	6.1	39.8	0.06 [2]	7.239	-.336	7.2	37.2
LMlsSt+MRBG	LM OLS est end Grd Status+ RIT Blk Mean Gth	0.78	5.551	-.047 *	6.9	46.1	-0.05 *	6.720 [2]	.036	7.2	38.6
LMeBSt+MRBG	LM EB est end Grd Status+ RIT Blk Mean Gth	0.78	5.444 [2]	.154	15.9 *	48.6 [2]	-0.05 *	6.817	.168	16.9 *	38.5
FLMls	Full LM OLS est end Grd Status & Slope	-0.31 *	6.882	-.295	14.8 [2]	36.9	0.11	7.936	-.313	11.3	32.7
FLMeB	Full LM EB est end Grd Status & Slope	-0.31 *	5.334 *	-.088 [2]	14.5	48.7 *	0.11	6.858	-.174	7.8	38.6
MnbZ	Mean obs norm-based means, sd to predict z	2.02	5.685	-.142	14.7	44.6	4.55	7.492	.074	13.3 [2]	28.8
MnbZ*	Mean obs norm-based means, sd (last double weighted) to predict z	1.97	5.560	-.192	5.5	45.1	3.90	7.074	.019 [2]	7.6	31.3

Grades 4-7 Fall and Spring Data Predicting Grade 8 Spring Status

		Reading					Mathematics				
MGLG	Obs end Grd Status + Mean Gth	1.26	6.280	-.372	5.6	36.9	0.06 [2]	6.772	-.095	14.1	42.2 *
MRBG	Obs end Grd Status + RIT Blk Mean Gth	1.54	5.913	-.154	11.8	43.0	-0.10	6.674 [2]	-.013 *	14.2 [2]	39.9
LMlsSlp	Obs end Grd Status + LM OLS est. Slope	-0.21 *	6.853	-.357	6.7	37.7	-0.47	7.447	-.344	9.8	36.0
LMeBSlp	Obs end Grd Status + LM EB est. Slope	-0.21 *	6.477	-.341	7.9	39.7	-0.47	7.160	-.328	7.6	38.1
LMlsSt+MRBG	LM OLS est end Grd Status+ RIT Blk Mean Gth	0.32 [2]	5.443	.055 *	9.1	47.7 [2]	0.32	6.608 *	.093	7.5	41.1 [2]
LMeBSt+MRBG	LM EB est end Grd Status+ RIT Blk Mean Gth	0.32 [2]	5.342 [2]	.180	12.1	49.9 *	0.32	6.687	.177	17.6 *	40.3
FLMls	Full LM OLS est end Grd Status & Slope	-1.42	6.339	-.196	12.3	41.2	-0.05 *	7.369	-.253	10.7	35.8
FLMeB	Full LM EB est end Grd Status & Slope	-1.42	5.301 *	-.063 [2]	13.6 [2]	47.2	-0.05 *	6.746	-.160	7.7	38.9
MnbZ	Mean obs norm-based means, sd to predict z	2.00	5.693	-.092	15.1 *	45.8	5.10	7.517	.063	10.7	26.5
MnbZ*	Mean obs norm-based means, sd (last double weighted) to predict z	1.98	5.578	-.125	5.8	46.1	4.67	7.243	.034 [2]	8.7	28.7

Residual plots.

Figures 1 through 4 present selected residual plots for Reading and Mathematics from the previous four sets of analyses (Cohorts A, B, C spring only, and C fall and spring). Each plot shows the resulting residuals from the selected model in relation to the final (observed) RIT scores. The plots selected for presentation were for the most parsimonious model in each set. For contrast and for illustrative purposes, the least parsimonious models for the same analysis set are presented in the lower portion of each figure. For purposes here, 'most parsimonious' refers to the model that resulted in the most favorable combination of low bias, low RMSE, low $r_{\hat{Y}resid.RITgi}$, and high percent of predictions within ±1 SEM.

The plots require little explanation but would benefit from pointing out a few characteristics of what we would expect to see in a parsimonious model. These include:

1. A trend line that runs through the range of the plot at or very close to the zero level. This is illustrated well in the Figure 2, Mathematics, most parsimonious plot.

2. When there is a positive or negative trend in the residuals, the difference between the most positive and most negative would be contained in a very narrow band. Figure 1, Reading, most parsimonious illustrates this.

3. Vertical scatter around the zero point would be compact, with the vast majority of residuals falling inside a narrow range (e.g., ± 10). Figure 2, for Mathematics, most parsimonious is the best example of this among the data sets.

4. Scatter around the zero point trend line would be vertically symmetrical across the entire measurement range of the RIT scores. None of the figures represents this particularly well, but Figure 2 for Mathematics, most parsimonious comes closest. Lack of symmetry is an indication that the model differentially accurate across the measurement scale.

Cohort A. The linear model with empirical Bayes estimates of end grade score plus RIT block growth was chosen as the most parsimonious for Reading. The plot for Reading (Figure 1) shows better predictions for scores above 200. More serious over-predictions (i.e., residuals < -10) were evident. For Mathematics, the linear model using ordinary least square estimates of grade status plus RIT block growth was selected. Again, the most discrepant residuals appeared at about RIT 225 and below..

Cohort B. Predicting Reading using linear least squares to estimate end grade and RIT block mean to estimate rate was selected as the most parsimonious model. (see Figure 2) This model resulted in a very slight bias toward under-prediction. Discrepant over-predictions (< -.10) were distracting but relatively infrequent. The empirical Bayes version of the same model was selected as the most parsimonious for Mathematics. Its pattern of residuals was fairly symmetric around zero and generally clustered within the –10 to +10 RIT range.

Cohort C, spring only data. The full linear model using empirical Bayes estimates was selected as the most parsimonious model for the Reading predictions. Even though this model resulted in a slight over-prediction bias (mean residual = -.31), its more severe under-predictions (residuals >10) were more common across the entire measurement range. This was similar to the most parsimonious model selected for Mathematics, the simple observed end grade plus RIT block growth model. Its more severe under-predictions occurred for scores in the 185-265 RIT range while its more severe over-predictions occurred in the 200-280 RIT range.

Cohort C, fall and spring data. The most parsimonious model for Reading was considered to be the linear model using ordinary least squares estimates for end grade status plus RIT block growth for a rate of growth estimate. The vast majority of its predictions fell within a 20 point band around zero. However, the severe over-predictions occurred for RIT scores in the 175-245 while the severe under-predictions were in the 190-255 RIT range of last observed scores. The model selected as the most parsimonious for Mathematics was the same as the one selected for the Cohort C, spring only data set. The comments made there apply to the fall and spring data set.

Figure 1. Residual plots of the most and least parsimonious models for Reading and Mathematics for Cohort A

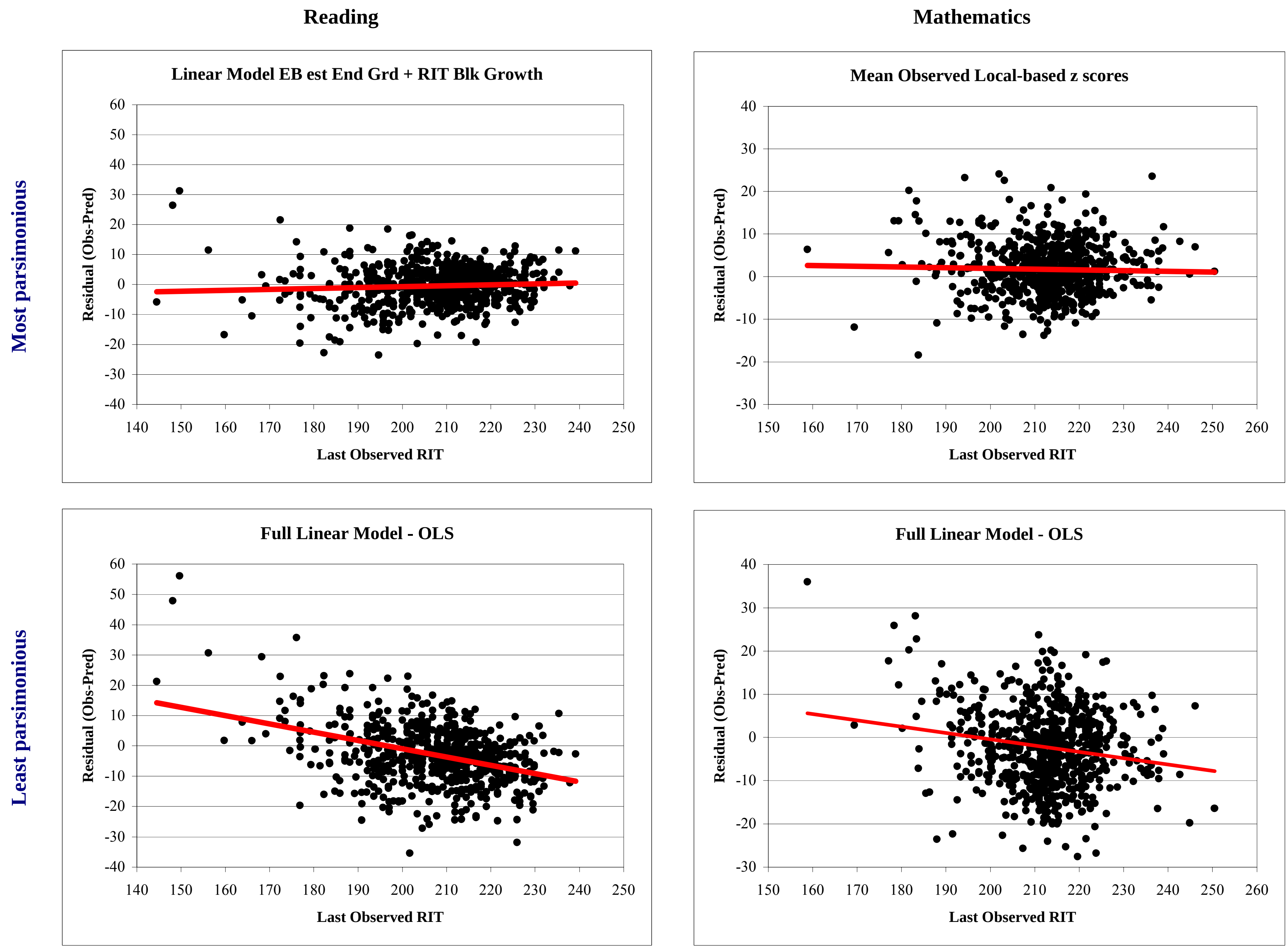


Figure 2. Residual plots of the most and least parsimonious models for Reading and Mathematics for Cohort B

Reading **Mathematics**

Most parsimonious

Least parsimonious

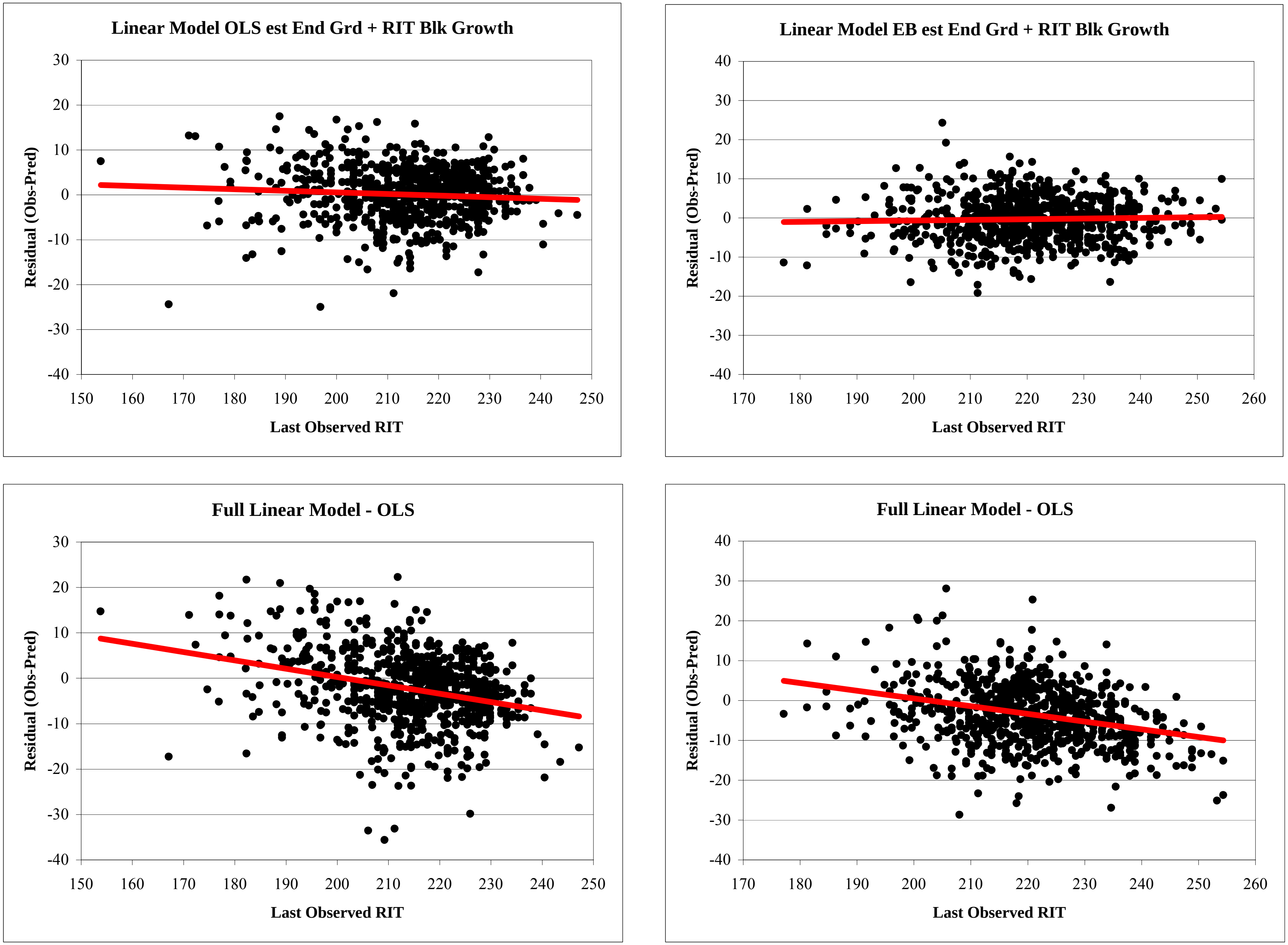


Figure 3. Residual plots of the most and least parsimonious models for Reading and Mathematics for Cohort C, Spring Data Only

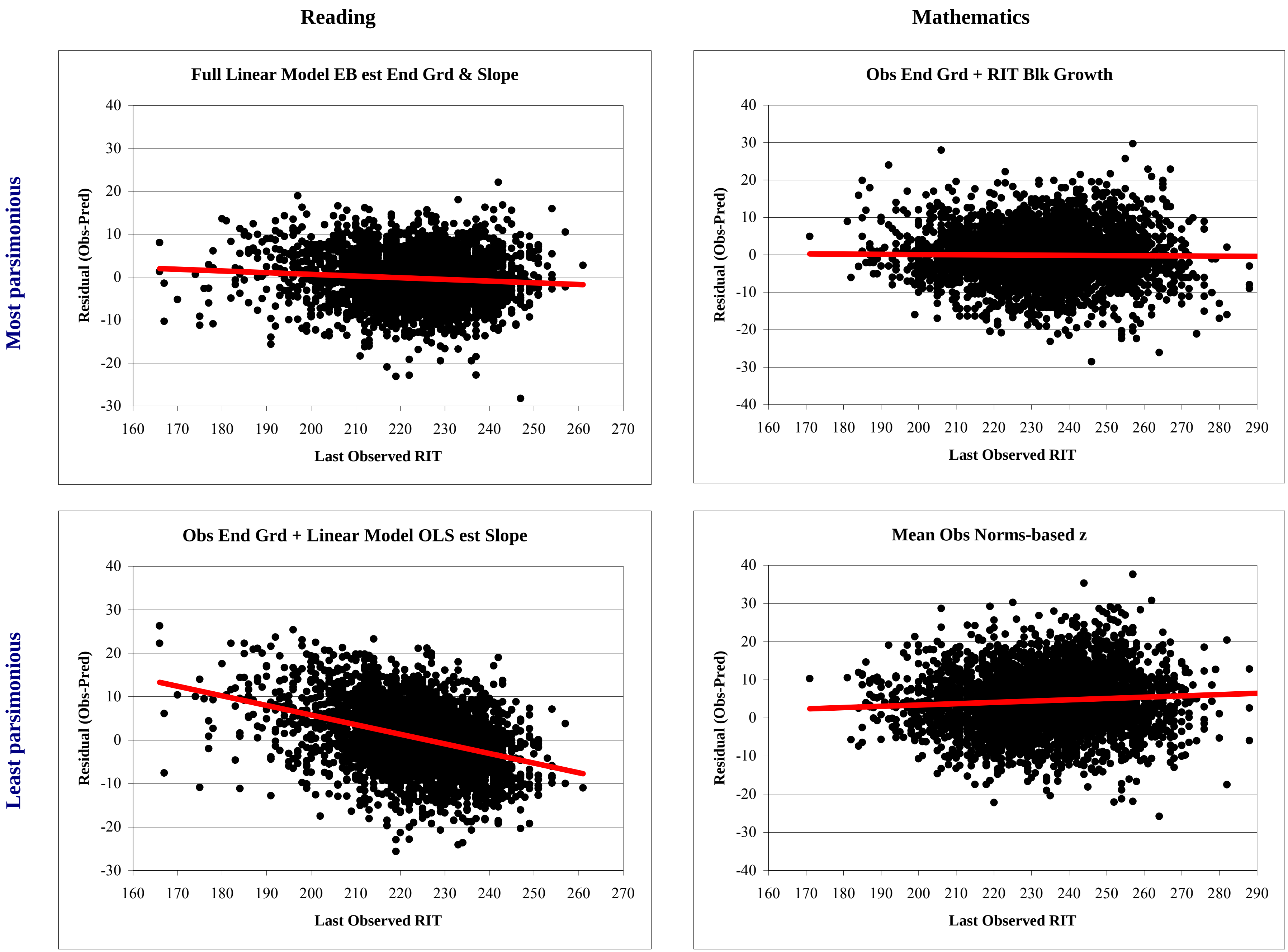
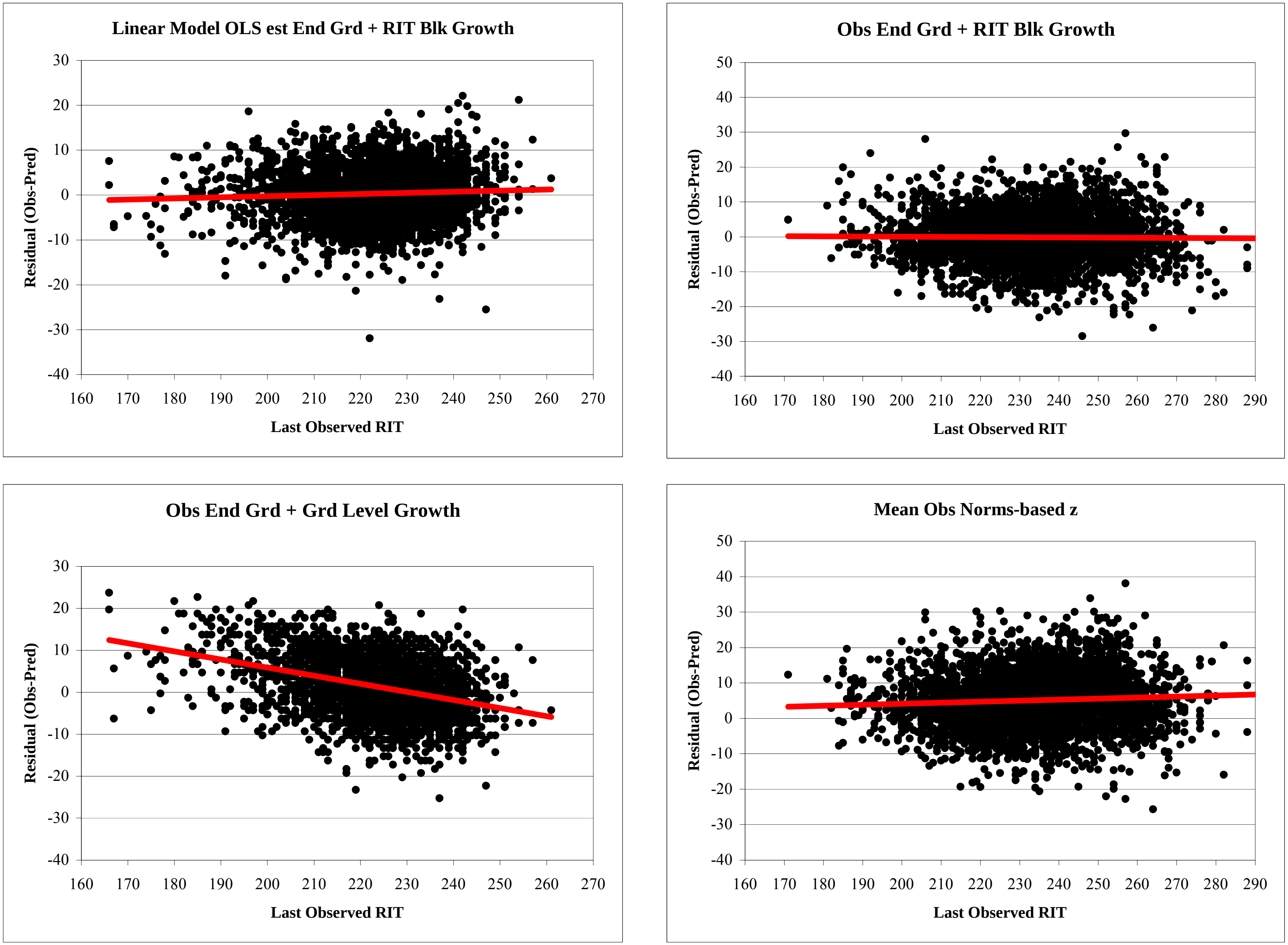


Figure 4. Residual plots of the most and least parsimonious models for Reading and Mathematics for Cohort C, Fall & Spring Data

Reading | **Mathematics**

Most parsimonious

Least parsimonious



Discussion

This study was undertaken to evaluate models that could be used to set single-year individual student academic growth targets. Multiple terms of individual student reading and mathematics test results were analyzed to predict each student's final status score in each subject. Test records from over 5300 students in three cohorts were used; two cohorts of roughly 670 to 750 students and one cohort of roughly 4000 students. The two smaller cohorts were from the same school district; the larger one was from the 2002 NWEA Norming Study and represented nine school districts. Three terms of spring data were used to predict scores in a fourth spring term for the two smaller cohorts. For the larger cohort, four terms of fall data and four terms of spring data were used to predict scores in a fifth spring term. Also, the four terms of spring data were used independently to predict scores in the fifth spring term.

The twelve models used to make predictions varied in the: a) treatment of data prior to the last 'observed' score, b) nature of the last score [observed or estimated], and c) estimate of rate of growth used [linear, RIT block growth, ignored in the *z*-score models]. Of the 12 models applied to each of the eight data sets, five emerged as yielding the most parsimonious set of predictions. The predictions within ±1 SEM of the observed scores ranged from roughly 40 to 50 percent for these models. Corresponding percentages for the six least parsimonious models ranged from roughly 11 to 37 percent.

The prediction task here was intentionally restricted to using only available achievement test data. Were a traditional modeling or forecasting approach taken, additional data such as school or district characteristics (e.g., class size, curricular differences), or student characteristics (e.g., gender, ethnicity, level of poverty, English language status) could have been added to help model additional variance. For example, recalling that Cohort C was made of data from nine school districts, it is quite possible that a good portion of the variability in the Mathematics data could have been attributable to differences in mathematics course taking patterns between these districts. Taking such differences into account, may have improved prediction accuracy. However, even though they may improve prediction accuracy, these variables would typically not be feasible to include. In all likelihood they would be viewed as setting differential growth targets (expectations) based on school and/or student characteristics; current collective thought cannot reconcile this practice with the demands of the standards movement.

In what might be considered a prophetic announcement of the results of this study, George E.P. Box (as cited in Sloane & Gorard, 2003), once opined, "All models are wrong, but some are useful." Even though the term 'parsimonious' has been used here to label particularly attractive sets of results for a model, the term could only be applied as a relative one. When the most parsimonious model accurately predicted (within 1 SEM) student status slightly less than 50 percent of the time, we can safely conclude that all these models are wrong, at least they are less accurate than we would like. However, this does not preclude the possibility that some of the models or model components may prove useful under specific conditions. What proves useful, may well depend on the characteristics of the data available to model. If a grade-independent scale can be assumed, the important characteristics for the models used here reduce to the quantity of data, the number of waves of data with common student test results, and variability in those data.

A district that has only one or two waves of same-student data, could in the absence of stable growth norms, assign individual growth targets based on the grade level differences in status norms. This is consistent with current standards-based accountability systems; all students in a grade would be assigned the same growth targets. Considering the potential disruption this could cause, it is not a recommended approach. A more promising approach would be to gather one or two additional waves of data and then investigate one of the two local-based mean *z*-score models used with Cohort A. These models should work well when the number of students per grade level is about 500 or more and the score distributions are approximately normal. When grade level growth norms are available, these could be used immediately, though for individual student growth targets they are only a slight improvement over using grade level differences in status norms. At the individual student level, growth norms that are segmented based on initial score (e.g., RIT block mean growth), will typically result in more reasonable growth targets.

When three or four waves of achievement data are available for making predictions, the range of options increases. Linear models that provide an estimate of the last (observed) score combined with a mean from segmented growth norms as a substitute for growth rate should be considered. Results from this study demonstrated that with short time series (e.g., 3 waves) the slope estimates of the linear models had low-moderate reliability (viz., .36 and .087 for Reading and Mathematics, respectively in Cohort A; and .12 and .31 for Reading and Mathematics, respectively in Cohort B). Use of the RIT block means in place of the rate estimates from the linear models, improved accuracy over the full linear models. In addition to the more complex models, the local-based mean *z*-score model could be explored when the conditions noted previously hold.

With four or more waves of data, consideration can be given to the full linear models. However, the results here demonstrate that more waves of data don't always yield the least biased or most accurate predictions, even though they are likely to be *among* the most accurate. A pattern of inconsistent term level variances for a cohort can be considered a sign that linear models, or at least the linear models used here, may not lead to the most accurate results (see Cohort C).

Explicitly including individual growth into a district or state level accountability system has the potential to expand the capability of the system by making it more comprehensive and more sensitive to the full range of academic change. To realize this potential, the expectations for academic change need to be generated from the perspective of the individual student. Research in this area is still immature and more research is clearly needed. However, even at this stage there is sufficient evidence to counter the unfortunate practice of declaring group growth targets in the absence of reasonable expectations for individual student growth – a practice that has been encouraged by status oriented accountability systems.

References

Flicek, M. & Lowham, J. (2001, April). *The use of longitudinal growth norms in monitoring school effectiveness and school improvement.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Kingsbury, G.G. (2000, April). *The metric is the measure: A procedure for measuring success for all students.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Linn, R.L., Baker, E.L. & Betebenner, D.L. (2002). Accountability system: Implications of requirements of the No Child Left behind Act of 2001. *Educational Researcher, 31,*(6), 3-16.

Raudenbush, S.W., Bryk, A.S., Cheong, Y.F., & Congdon, R.T. (2001). *HLM 5:Hierarchical linear and nonlinear modeling.* Lincolnwood, IL: Scientific Software International.

Seltzer, M., Choi, K., & Thum, Y.M. (2002a). *Latent variable modeling in the hierarchical modeling framework: Exploring initial status x treatment interactions in longitudinal studies.* (CSE Technical Report 559). Los Angeles, CA: Center for the Study of Evaluation, Standards and Student Testing.

Seltzer, M., Choi, K., & Thum, Y.M. (2002b). *Examining relationships between where students start and how rapidly they progress: Implications for constructing indicators that help illuminate the distribution of achievement within schools.* (CSE Technical Report 560). Los Angeles, CA: Center for the Study of Evaluation, Standards and Student Testing.

Sloane, F.C. & Gorard, S. (2003). Exploring modeling aspects of design experiment. *Educational Researcher, 31*(1), 29-31.