Measuring Assessment Results Over Years

May 31, 2018

0.1 Measuring Assessment Growth

A key question coming from a deeper understanding of standard errors of measurement are the consequences on trying to measure the academic growth for a single student and the entire program. From our prior reviews of the literature, we found that the SEM is considerable, typically as large or larger than the expected growth for the student within a single year. The implication of this amount of error, is that we worry that any attempt to model or measure growth, particularly within a single year, might be far too influenced by error to yield reliable numbers.

In order to concretely explore this scenario, we make a measurement of the consequences of SEM by applying a Monte Carlo sampler. After running simulations of our program within a single year, over 2 and 3 years, and even an extended 10 year simulation we have been able to come some determinations about our current ability to measure student improvement at the site, school, and student level. Results are summarized as follows:

Heirarchy Level	Measuring Improvement under SEM
Site	Within a single year the average student improvement is relatively close to the actual student improvement. We can reliably use the number without thinking too hard about measurement error.
School	Measurements at the school level will always be noisier than those measured at the site. We have to be more conscious this natural volatility and not penalize/reward PMs for outcomes caused by error
Student	We are completely screwed. Even over 10 years we couldn't get an accurate measurement of their improvement (that is get less than 80 point standard deviation)

0.2 Findings

As summarized above, our findings have implications for City Year's ability to measure assessment improvement accurately, particular at the school and student level.

School Findings In these simulations we observed that the considerable noise that merely reducing the sample size adds when measuring the average student improvement for a school. When simulating this, we avoided the question of growth goals, because of the computational complexity that it would have introduced in the sampler. However, it is undoubtedly fair to say that the

proportion of students who meet their growth goal is directly propotional to average improvements of students at that school, and therefore if there is considerable variance in the average improvement than the variance would translate into those those meeting the growth goal as well.

These circumstances imply that we need to apply special care when looking at the percentages of students who meet the growth goal for students, especially when trying to evaluate if a school was below or above average. A school will need to deviate significantly (though what is technically significant isn't 100% clear) for us to actually believe that the deviation is caused by anything beyond noise.

Student Finding Our student findings paint a rather bleak picture of our hope that we'll be able to some day measure student improvement. We ran simulations of students over 10 years and even then if we apply a naive approach to measuring improvement (post - pre) then we will face huge measurement error (+/- 150 points) between the true improvement and the measured one.

In order to get closer to a true measurement of student improvement, not only do I believe that we would need to look at year over year measurements but I believe we would need a new method for combining multiple HMH measurements into a single number *with lower SEM any individual MI/RI score has.* By replacing pre and post values with aggregated values with less SEM, the probability distribution for post - pre will have less variance and (hopefully) give us a much better measurement of improvement. This could then get aggregated up into the school and site level. Theoretically it shouldn't differ too substantually from the naive method at the site level.

We haven't actually developed the method, but a sketch of it might be Bayesian where we start with a prior understanding of that grade levels distribution of student abilitiy, and is then subsequently updated with each score. Even after 2 scores we should be able to provide a third aggregate score with less SEM (assuming the 2 scores roughly agree).

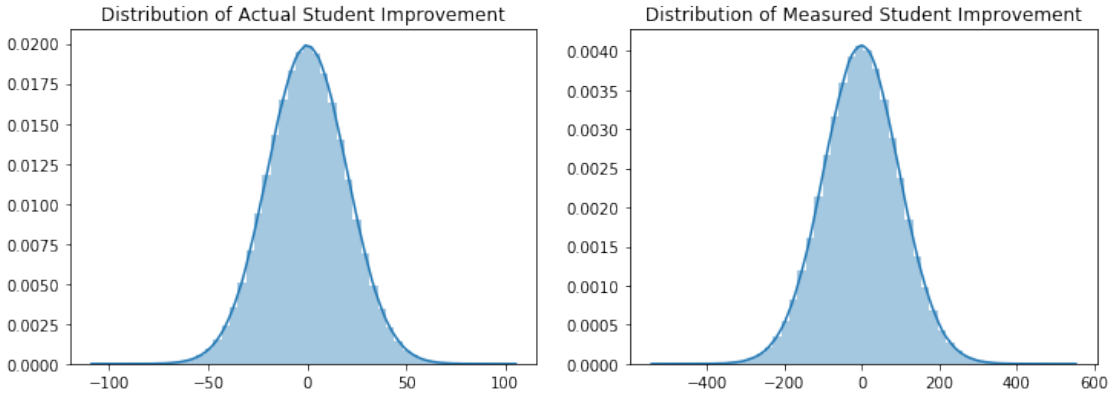
0.3 Analyzing Simulations

Our goal is to improve our understanding of the impact of SEM on our ability to measure true student improvement. This approach entails us essentially simulating a group of 3000 students, providing them "true" baselines, and then simulating the successive sampling of HMH scores over several windows. By doing this we can simulate a thousand or more program years at once, and observe the various possible outcomes that might occur.

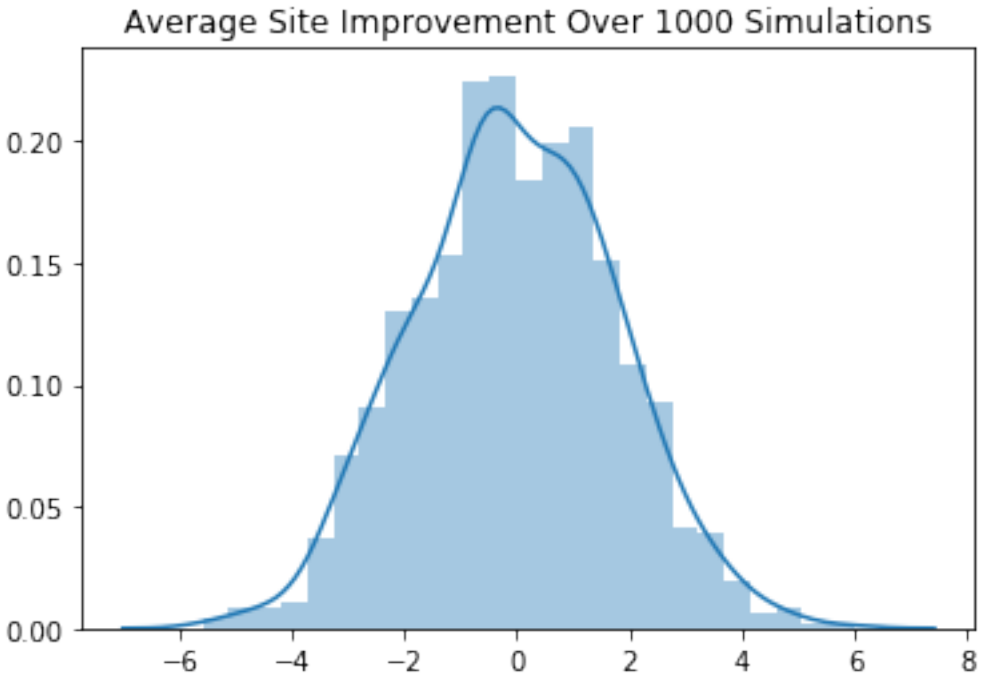
To be certain, this simulation is by no means a perfect representation of CYLA's program. However, we have built into the simulation many of the same assumptions that we now believe to be true of our real-life students, lending some credence to the usefulness of these simulations. Any improvement in our understanding of our real-life student's situations will of course also improve our ability to simulate and model the possibilities.

0.3.1 What does it look like if our student's aren't improving on average?

In this first set of simulations we consider a set of students who generally aren't improving. We could simulate this by literally having their true value not improve at all, but just to be a bit more realistic we're drawing from a normal distribution centered at 0, specifically $\mathcal{N}(0, 10)$. To translate, this is simulation of 3000 students who on average aren't improving over time, though their ability isn't static year to year.



One of the nice things about this simulation approach is that we actually know what the true abilities of our simulated students are, so we can observed the difference between the true values and the ones we measure through SEM. For example, pictured above is the measured improvement for our 3000 students compared to the actual values. Notice how the distribution on the right is a much larger range, stretching +/- 200, as opposed to the true values which don't go much farther than +/- 50. This highlights the real danger in trying to over utilize the pre/post or improvement values for a single student, since it is so heavily influenced by the measurement error.



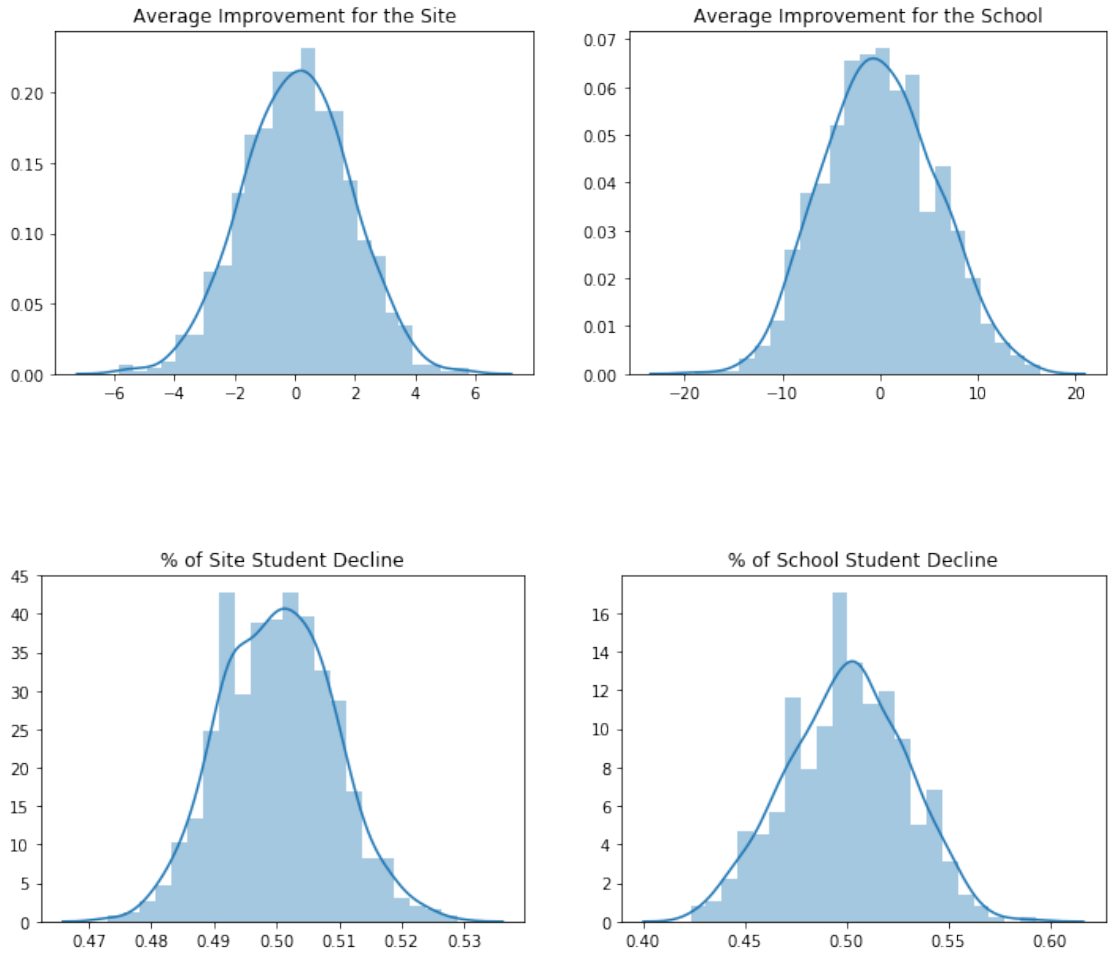
On the other hand, as we aggregate up to the whole program we are relieved to see that the average improvement, falling in between -4 and 4 in our simulations, nicely reflected the true

value. This reasuringly tells us that at 3000 students we can be relatively certain that the average improvement of the sample is close to the true average improvement.

So we've discussed and n=1 and n=3000 cases, but what about inbetween.

0.4 The difference between measuring a single school vs the site

Consider below where we run simulations on populations of students closer to an individual school team (somewhere near 125) compared to analyzing the entire site.



What we should immediately notice is that by reducing the number of records by a power of ten, we start to observe greater differences between runs in the Monte Carlo. For example, when we measure out program as a whole with 3000 students we find that the average improvement over all is tightely bound around 0, spread predominatenly between -4 and 4. With a school size population of 300, the range increases by quite a bit, out to closer to -15 and 15.

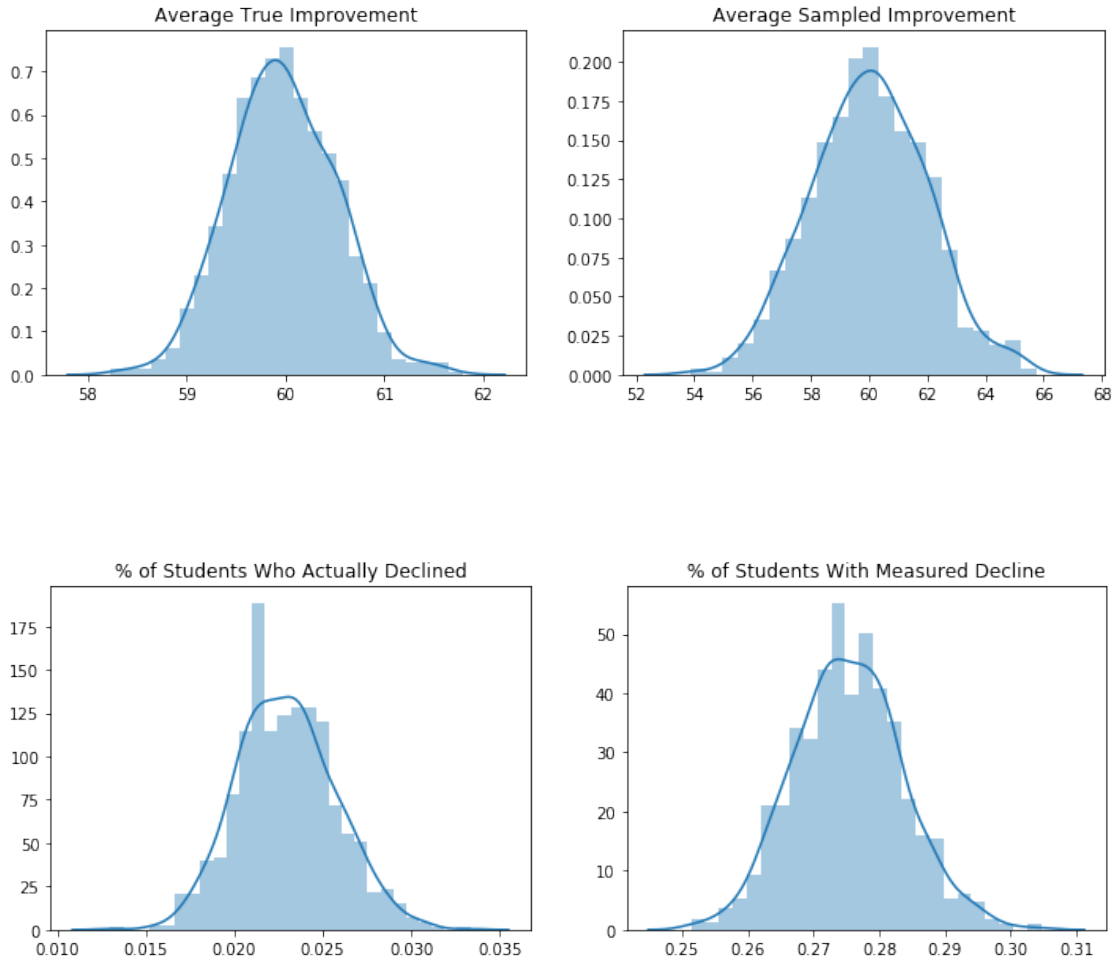
0.4.1 What challenges does this increases in variance cause?

Well, we regularly measure these values and related values at the school level and have a inherent bias to consider PMs whose students improve more on average as somehow more effective. What

we should notice though is that in our simulations we see many examples of samples which differ by more than 30 points, when we know for certain that each student is, on average, not improving at all. Therefore this 30 point different is *caused purely by noise.* Likewise, some schools will see 40% of their student decline, and some will see 60%. Again, not because there is any meaningful difference between the schools. Smaller schools (elementary schools) will be even more prone to this effect, and it's likely that we'll see the most extreme variants from these schools, whereas we'll see values closer to the true mean for larger schools.

0.5 Adding Real Improvement

In the next simulation, we simulate a population of students who are improving on average, though the distribution allows for a small probability that their true ability decreases with a period.

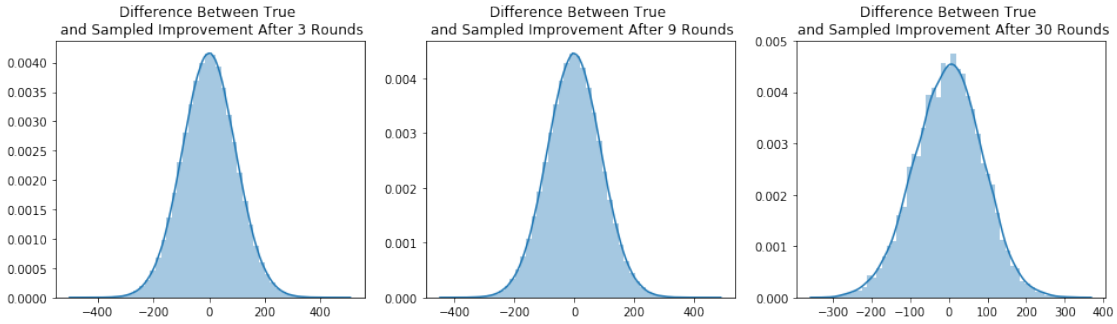


Notice how within a single year we know that only a very small percentage of students will actually have a declining lexile (between 1 and 3 percent), but as a result of the SEM we'll actually observe a large percentage of students with measured decline (25-30%).

0.6 The Discreancy Between the Measured and True Improvement

We would love a method which allows us to get a better understanding of the true improvement of single student, since so far it seems like we have a hard time even talking about school level results without being incredibly mindful of the measurement error.

One hope we had was what if we measure results year of year we might be able to smooth over the the error and give us a better idea of the true improvement. In the next set of simulations we compare how our program would look if we measured the same set of students over 3 and 10 years.

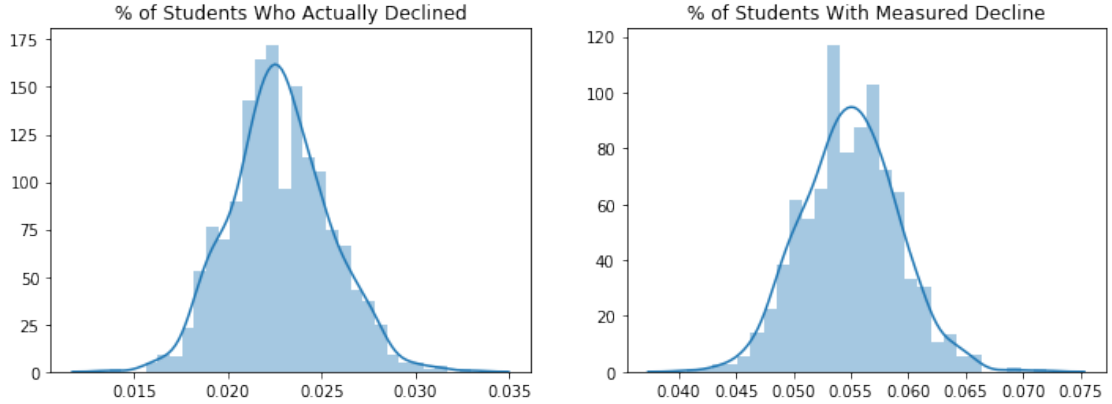


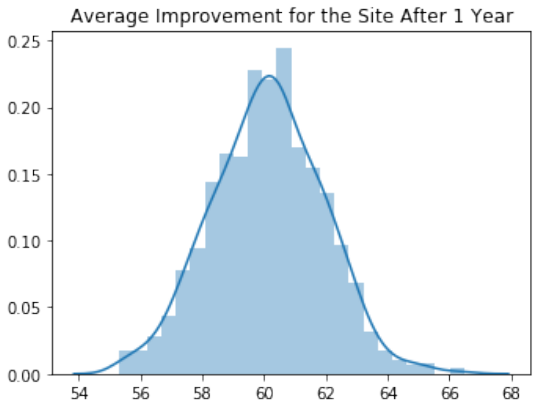
0.6.1 Bad News for Single Student Improvement

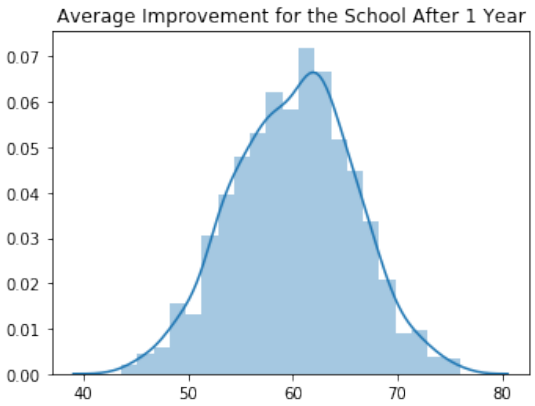
We see in the pair of histograms above that the actual reduction in difference between a students measured and actual improvement is still quite small after 9 and even 30 rounds of assessment. The bad news is that this suggests we're extrememly limited in what we're able to say about a single students improvement even after many assessments. A significant proportion of students will have their measured improvement differ significantly (by 100+ points in either direction) from the true value.

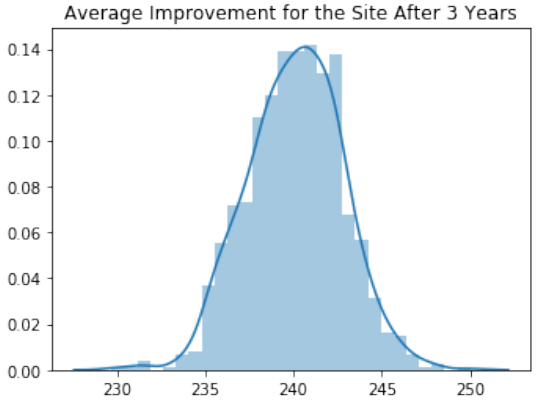
0.6.2 Good News for Student Decline

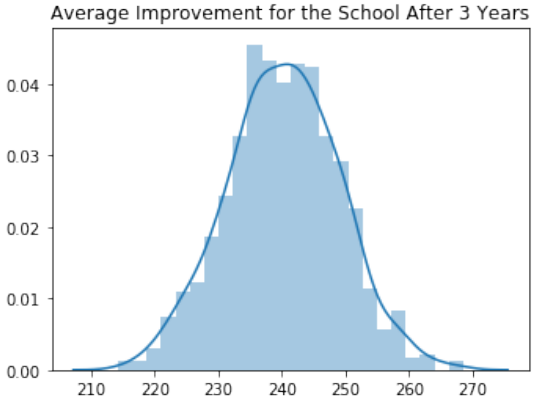
However, multiple years of measurement does significantly reduce the number of students who we measure as declining which converges towards the actual value.




Average Improvement for the Site After 1 Year
0.25
0.20
0.15
0.10
0.05
0.00
54
56
58
60
62
64
66
68


Average Improvement for the School After 1 Year
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
40
50
60
70
80


Average Improvement for the Site After 3 Years
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
230
235
240
245
250


Average Improvement for the School After 3 Years
0.04
0.03
0.02
0.01
0.00
210
220
230
240
250
260
270

In [ ]: