Measurement Error in the MI and RI

May 31, 2018

Abstract

The HMH Reading and Math Inventories (RI and MI respectively) are key tools in the way that City Year Los Angeles (CYLA) seeks to deliver its service model. Despite giving us greater clarity on our students academic ability, the data also introduces its own sources of confusion. **A third of CYLA students typically show a decline in ability from first to second administrations of these assessments, and a significant portion (about 22% overall) continue to show measured decline in the third administration.** These measurements defy the common belief that a students true academic ability won't decline at all within an academic year. In this paper we review prior research published by Scholastic and HMH into the measurement error inherent in their Math and Reading inventories and from this build out a model to visualize what the potential influence of measurement error is for a single student. This model shows that the likelihood of measuring decline is directly proportional to the amount of true improvement that a student experiences. Based on these results, we then estimate how many students would have measured decline based on the observed results in the 2016-2017 program year. These estimates indicates that a large proprotion, though not all, of the decline we observe does come from measurement error. The remainder of the error is presumed to be a result of other factors such as student motivation or the testing environment, but are not directly measured.

1 Introduction

In the 4 years that City Year Los Angeles (CYLA) has utilized the HMH Math Inventory (MI) and Reading Inventory (RI), it has consistantly been observed that about a third of assessed students will decline from the first to the second assessment. Over the years there has been considerable conversation on what might be the cause of this decline for a seemingly large proportion of served students, and many hypotheses have been proposed related to ACM investment in assessments, student motivation, and variations in environmental conditions and student conditions.

However, recent reviews of research conducted by HMH on the MI and RI have begun to offer some clarity on an additional source of measurement error for the MI and RI, the measurement error inherent in the assessment itself. This information in turn has enabled us to begin to model what our expectations for our students might be.

In this paper we will start by seeking to understand the research on the measurement error inherent in both assessments. We will then consider the implications for a single student and quantify the likelihood that we observe a decline for that student. From this intial model we will then estimate the number of students we would anticipate measuring a decline, based on the data availabe from 2016-2017, and compare that to the actual results.

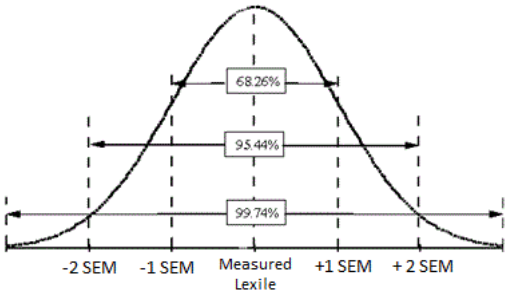
2 Measurement Error in the MI/RI

Intangible concepts like "reading comprehension" or "mathematical understanding" are hard to measure. While ideally we would measure reading comprehension like we would a student's height, realistically we have to find indirect methods to represent and measure a student's reading comprehension or math ability. For most math or reading assessments, the strategy is to try to correlate a student's ability to answer a small number of questions, with a corresponding measure of ability (e.g. the Lexile or Quantile). However this strategy is imperfect and is prone to error. First, there are human factors that cause variation in a students performance on the test. This includes things like whether they managed to get a good night of sleep, ate breakfast, or had a distracting testing environment. This difference between a good day and a bad day for a particular could certainly be a signficant different from these factors alone.

The other major souce of error is the inherent error in the model caused by the assessment's reliability to measure reading/math ability. This source of error is typically called the standard errors of measurement (SEM) While a student's ability to answer questions certainly is related to their ability, there are many potential sources of error in this strategy. For example, a student doesn't know every math topic evenly. So based on the questions that a student is given, they may do particularly well or poorly based on if they new many of the questions or not. Likewise in the RI, students may be more familiar with some words than others, which will effect their ability to get those questions correct. Based on research from Hambleton et al. (1991), the RI and MI use several strategies to minimize the SEM including ensuring that the assessment is long enough to ensure a good sampling of different questions, having highly discriminating questions for which answers can't be guessed, and dynamically adjusting the difficulty of the test to try to deliver questions which are as close to the students true ability as possible.

The SEM can be visually displayed with a normal "bell" curve as we see in figure 1. Suppose we give a student the RI and it spits out a 500L as their Lexile. The SEM tells us that 500 is probably not their "true" reading ability (their Lexile if our assessment was perfect), but that their true ability is in the neighborhood of the measured Lexile. The most likely place for the students true ability is within 1 SEM of what we measured. So, for example, if the SEM of the RI is 50 then the most likely place we'll find their true ability is in the range 450 to 550. As noted in figure 1, the probability is actually about 68% that the true ability is in that range. As we expand the range, we can capture a greater percentage of likely scores. So we can be 95% confident that their true ability is in the range 400 to 600, or 2 SEMs from their measured score.

Figure 1



2.1 Standard Error in Measurement for MI & RI

In a report published by HMH titled *Accuracy Matters: Reducing Measurement Error by Targeted HMH Reading Inventory Testing*, the SEM for the RI is thoroughly explored. One import point they raise is that because the RI & MI are adaptive tests, the SEM for each student is actually unique and varies at each assessment. Therefore as we discuss SEM keep in mind that we have some uncertainty over precisely what these values are for each student. However, HMH provides sufficient guidance that we can reasonable approximate the SEM for a student in several situations and go on to make estimations about the consequences.

The most important factor in determining what the SEM might be for a particular student is the number of questions a student has answered. In *Accuracy Matters* and the technical guides, they explain that the assessment is able to perform two main types of differentiation. Once a student has answered 40 questions in the assessment, which is equivalent to approximately two administrations, the assessment is able to differentiate the questions it asks based on both their grade level as well as the established reading or math ability of the particular student. However, prior to reaching those 40 questions the assessment doesn't consider itself reliable enough to differentiate based on ability, and so only differentiate the questions by grade level. This leads to essentially two different sets of SEMs based on if the student falls into the first class or the second. Extrapolating from *Accuracy Matters* we determined reasonable estimates for the the SEM in each round of the RI:

Round	SEM w/o Established Reading Level	SEM w/ Established Reading level
1	91L	56L
2	91L	56L
3	56L	56L

For the MI HMH has done less in detailing how the SEM will decrease as the system collects more information on a particular students. In their Technical Guide HMH identified the SEM for MI as being 70Q and in a more recent guide on interpreting student declines adjusted the estimate down to 63Q. However, again that SEM is predicated on the student having an established baseline within the system, and thus prior to taking two administrations we should expect the SEM to be higher. While we don't have precise numbers, it will be usefull to make esimtates of these values in a table like we have for the RI.

Round	SEM w/o Established Math Level	SEM w/ Established Math level
1	100Q	63Q
2	100Q	63Q
3	63Q	63Q

While for either assessment the SEM will vary from student to student, we will use the values from above in conducting the modeling as they function as a reasonable approximation for setting our expectations in response to measurement error.

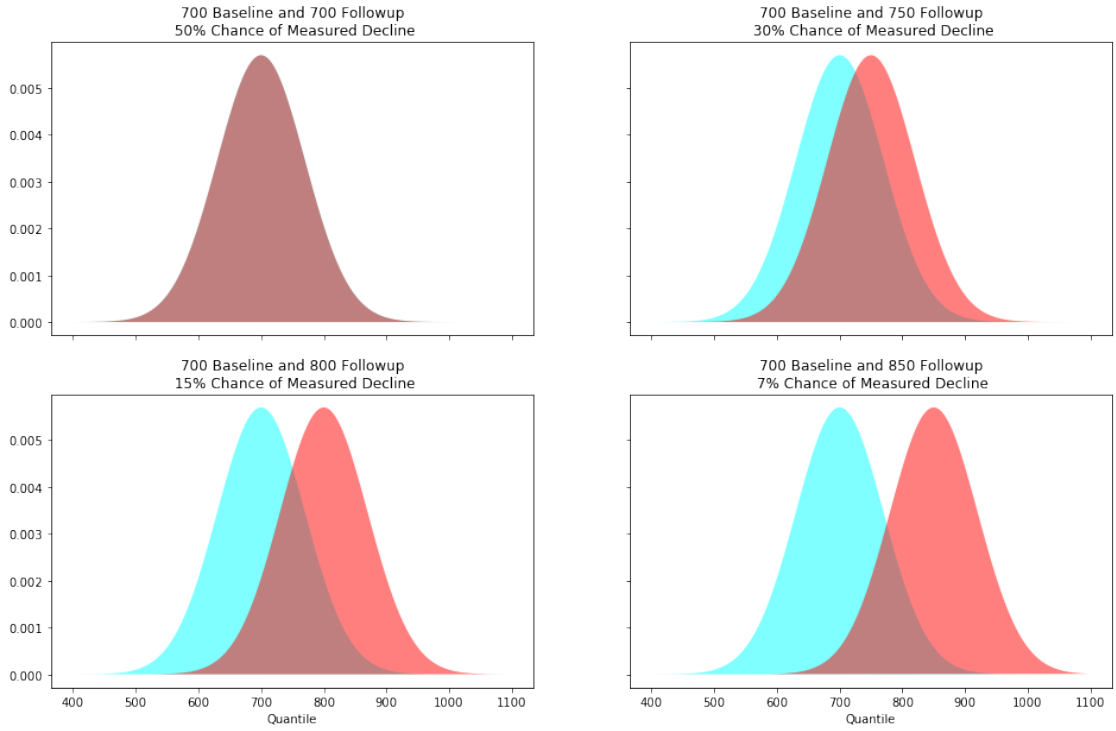
2.2 Decline for a Single Student

Let us consider the consequences of SEM for a single student as they take several assessements. It is helfpul to actually think about the consequences by SEM, particularly in terms of trying to

understand the likelihood of student decline, by considering what might happen if we knew a students "true" ability.

Suppose that we know that a students' true ability at the baseline is 700Q and that between the baseline and their second assessment they improved in true ability by 50 points. How might this situation play out when measured by HMH? Examine the plot in the top right corner of figure 2 titled "700 Baseline to 750 Followup" to see the graph corresponding to this situation. Notice that this graph has two distributions which are centered at 700 and 750, and which represent a possible range where HMH would measure their first and second quantiles. So, to be explicit, this graph suggests that one possibility is that we could assess this student twice and get a 750 for their baseline and get a 850 followup. However it also shows that extreme cases are possible, such as a 850 baseline and a 600 followup. When we consider all possibilities for this student, we can calculate that there is a 30% chance that a student in this situtation would have a measured baseline and followup that corresponds to a decline. In figure 2 we show the distributions for the baseline and follow up measurements given four different situations where their true math ability in the follow up is 700, 750, 800, and 850. Key to understanding figure 2 is seeing the correlation between the probability of measuring a decline and the overlap between the two distributions. When the distributions are perfectly overlaid then the probability of a measured decline is just a flip of a coin, 50%. However, as a student's improvement grows the distance between the two neighborhoods grows as well and the overlap between the two regions decreases in size.

Figure 2



2.3 Main Factors in the Probability of Decline

From our initial review of data and analysis of a single student's outcome, we conclude that two main factors determine the probability that a student will have a measured decline from baseline to follow up.

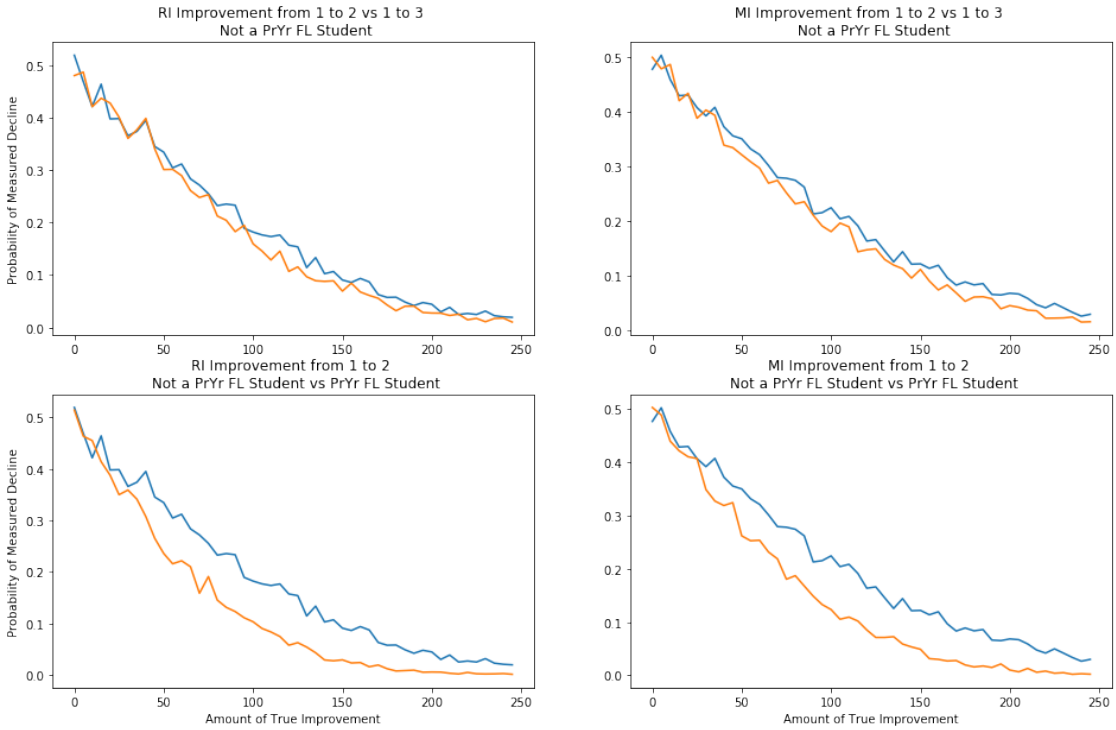
1. **Measurement Error** - As HMH pointed out, the SEM declines as a student gets assessed more. It also drops significantly after a student has been assessed twice in the system. As the SEM declines probability of measured decline will also decrease.
2. **Amount of True Improvement** - The probability of a measured decline reduces as the amount of true improvement increases.

To build a deeper understanding of these factors we produced figure 3. In these 4 visualizations, we consider comparisons which demonstrate these two points. Generally, each figure shows a graph of the relationship between the amount of (true) improvement that a student experienced between the two assessments, and then graphs the likelihood that we measure a decline for that student based on our analysis. Key to notice in each graph is that the liklihood always starts at about 50% if a students didn't actually improve at all, and then as the student improves we see the likelihood of a measured decline decrease. Each of these visuals will nicely demonstrate point number two from above, but to demonstrate point number 1 we have used these visuals to do several comparisons.

In the top row of visuals, we compare the likelihood of student decline in the measured improvement from the first to second administration to the first to third administration for both assessments. Key to notice is that the orange line, which corresponds to the improvement from the first to third assessment, is lower than the blue for both assessments. This is a result of the assessment being able to differentiate by the individual student's ability, instead of simply by grade level as in the first two assessment. The reduced SEM causes the likelihood to trend to zero quicker.

The bottom row is a comparison of students who have established baselines in SAM at the start of year to those who don't. Key to notice is that if we have an established baseline for that student, the likely of measuring a decline in their assessments drop considerably quicker. For example, we can notice that wereas we have a 20% chance of a measured decline in the MI of a student without an established baseline (i.e. not a prior year focus list student), the student with the established baseline will have closer to a 12% chance of a measured decline.

Figure 3
Improvement vs Likelihood of Measured Decline



3 Key Findings

3.1 Finding #1: Measuring a decline is possible for *every* student

What we can see clearly in the analysis of a single student is that a decline is possible for nearly any student. Even for a student whose reading ability improved by 100 points, there is still a ~18% chance of them measuring a decline from the measurement error alone. If we consider the other potential factors that might influence a students test scores including the many environmental, emotional, and social reasons why a student may or may not perform their best on an assessment, then we should consider that this probability is likely even higher for any given student.

3.2 Finding #2: But it's more likely in the 2nd Administration than in the 3rd

The probability that we measure a decline in a students pre and post assessment for the RI and MI is proportional to how much they truly improved. **Since students will necessarily have grown less by the second administration compared to the third, we therefore expect to see more decline in the second Administration.**

3.3 Finding #3: SEM explains much, but not all of the decline we observe

As a model for what this process might look like, we'll examine FY17 data. For each MI/RI round 2 and 3, we went into the data and calculated the % declines that we actually observed. We

also calculated the average improvement from the first to latter assessment and by using this as an approximate for how much true improvement our students experienced on average we consider what % decline the SEM would explain. We should expect to see that the actual decline is somewhat higher than the predicted decline, since SEM only accounts for a portion of variation in a student's measure Lexile/Quantile. Finally, we should note that in calculating these values we made several estimatation which will add error to these predictions. In particular we are considering the average improvement as a proxy of their true improvement and an average SEM instead of the specific SEM for each student, since neither the true improvement nor the true SEM is a known for any student.

Assessment/Round	Average Measured Improvement	% Decline	Predicted Decline (based on SAM documentation)
RI Rnd 1 to 2	50.7	29.5%	26.5-33.8%
MI Rnd 1 to 2	49.8	36.2%	27.7-35.0%
RI Rnd 1 to 3	106.8	21.9%	8.9-15.3%
MI Rnd 1 to 3	132.2	23.7%	6.7-14.2%

Key to notice in this table is that our predicted decline based on SEM is lower in round 1 to 3, but the actual decline tends to be 6-8% higher. A portion of this difference may be because a higher proportion of our students still only have two assessment scores in the find round and thus have a higher SEM. There also may be more error introduced from distracting environemntal conditions, testing fatigue from many standardized tests during the end of the year, or the general excitement and distraction of being nearly to the end of the year.

4 Conclusion and Recommendations

The consequences of these findings extend all the way to the AmeriCorps member, nearly all of who will experience having a student decline as a result of the SEM. We therefore should consider how this influences how we talk about and utilize the RI and MI in practice.

4.1 How we talk about Lexile and Quantile

To fully appreciate the consequences of SEM on RI and MI results, a level of fluency is statistics is required which makes this concept less approachable for ACMs and program staff. However one way to frame conversations about Lexile and Quantile scores, and in particular their decline, while keeping in mind the SEM might sound like this.

"Lexile and Quantile scores are best thought of as indicators of the neighborhood of a students ELA/Math ability. That is, when the assessment tells us our students ability is a 563, the specific number that the student got isn't actually what's important or what's helpful. That same student, on a different day, could score a 500 or a 600. On a really bad day, they might even score a 400. So when we're preparing interventions for this student, we want to consider content that is appropriate within a wide range around given number. We can improve our decision making here, and choose more appropriate course material by then incorporating other pieces of data, like the classroom content and your specific knowledge about the student. In this way, the score really serves *as a starting point* from which we investigate further, and not as a silver bullet which tells us precisely what to do."

4.2 The importance of keeping prior year focus list students

HMH was very clear that one of the easiest way to reduce SEM was to keep the student in the system year over year. The more times the student is assessed, the more the SEM will be reduced and the more accurate our pre and post scores will be. Therefore we can help to reduce the amount of decline we see by keeping a greater percentage of students who have taken the HMH assessment year over year.

4.3 Accounting for SEM in Future Research

When performing formal statistical analysis of student's MI and RI score we may be better able to explain the variance in scores we see by including SEM in the models. One method which would accomplish this would be to develop a new metric which measures the probability that a student improved by some threshold. To do this we consider the probability distribution of the difference between the distributions representing pre and post scores. An example of this metric might sound like: a student has a measured baseline score of 674 and a follow up of 737. With this information and the SEMs for their pre and post, we can calcualte that the probability that this student improved by at least 0 points is 71%. We can extend this idea so calculate the probability that a student improved by any given threshold. So, suppose we want to know how confident we are that a student improved by at least their growth goal. If the growth goal was 75 points, the probability that their true improvement was greater than the growth goal would be 46%.

4.4 Retests

The current policy allows for retesting a student who measures a decline by 100 points or more. However, depending on the situation this can be quite a common event. For example, if a student was taking the MI for the first time then the SEM is 100 points all by itself. A quick glance at figure 3 and we can determine that even if all of our students taking the MI were to have improved in truth by 50 points, we would see approximately a third of them decline. A retest policy which took into account SEM might be that a student would need to decline by at least 2 SEM to consider their score sufficient for retesting. This shift would add some burdon on the impact analytics team to keep track of if a student has a baseline established in SAM at the beginning of the year and then dynamically calculate if a student would qualify for a retest based on a decline.

4.5 What if the baseline was inaccurate?

In the case that a student sees steep decline on the RI, say from 700 to 200 Lexile, then we would be inclined to believe that the followup score was in error and that the student should be retested. However, we also see in the data that students will go from 200 to 700 Lexile, which likewise indicates that the baseline was in error. Based on new understanding of the RI and MI, what might be a reasonable policy? One potential option is that if we determine that the first score was likely in error, then we might instead calculate their growth from the second score. Concretely, if a student tested in all three rounds with the following scores: 200, 700, 750. Then in this case we might calculate their improvement from the second to the third assessment.

4.6 Measuring Improvement on HMH Assessments with SEM

In subsequent research we examined the impact of SEM on our ability to measure student improvement, which clarified some questions we had about how effectively we could make mea-

surements of student improvement at the individual school and student level. In summary, at the site level (roughly a thousand students) we can make accurate measurements of average improvement for the entire site. Thankfully, the law of large numbers will mitigate the error caused by SEM. However, at smaller scales such as the school level (closer to a hundred students), and of course at the individual student level, we find that there is considerable error in the distributions of student improvement. It is sufficient that for any particular student our confidence interval would necessarily be a very wide range (+/- 200 points for example). For a particular school

5 References

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications, Inc.

HMH. (2016) *Accuracy Matters: Reducing Measurement Error by Targeted HMH Reading Inventory Testing.*

HMH. (2014) *Math Inventory Technical Guide.*

HMH. (2016) *Interpreting Assessment Results.*

Scholastic. (2007) *Scholastic Reading Inventory Technical Guide.* Scholastic Inc.

6 Appendix A

Below we demonstrate the probability calculation discussed *Accounting for SEM in Future Research* above.

Out[33]:

	RI Baseline	RI Follow Up 2	P(RI Imp 1 to 3 > 0)
2	3.0	363.0	0.998985
3	674.0	737.0	0.705412
4	476.0	660.0	0.942629
10	476.0	456.0	0.431940
11	159.0	384.0	0.973115