REL MID-ATLANTIC Regional Educational Laboratory at Mathematica

# Using Existing Data to Create New Diagnostic Measures for States and Districts Webinar Transcript

**[Webinar Producer]:** Hello everyone. Thank you for attending today's webinar. Before we begin, we would like to go over a few housekeeping items. At the bottom of your audience console there are multiple application widgets you can use. You can expand each widget by clicking on the Maximize icon at the top of the widget or by dragging the bottom right corner of the widget panel.

A copy of today's slide deck and other resources are available in the Resource List widget indicated by a green file icon at the bottom of your screen. If you have questions during this event, please click on the purple Q&A widget at the bottom to submit your questions. We will take questions throughout the webinar.

If you have any technical difficulties, please click on the Help widget. The question mark icon covers technical issues. You can also press Command R or Mac F5 to refresh your player console. You may also use the Q&A widget to submit technical issues. Finally, an on- demand version of this web cast is available one day after the webcast in the same audience link you used to join today's events. Now I'd like to introduce Brian Gill. Brian, you now have the floor.

**[Brian Gill]:** Great. Thanks so much, Brian. Thanks. I want to thank everyone for joining us today. It looks like we've got a good group coming on. And I have to say, I particularly appreciate your participation at a time when so many of us have had our work and our lives disrupted by Covid-19. Of course, when we started preparing this seminar, we couldn't have anticipated conducting it as a time when the vast majority of school buildings in America would be closed, leaving 55 million students at home in this country, and many, many millions more around the world. But the school buildings will eventually be reopened, and even while they're closed, educators and policymakers, as you know, are working hard to make sure that students keep learning.

The ten regional laboratories that are funded by the U.S. Department of Education, including the Mid-Atlantic Lab, which produced this work that you're going to hear about today, all of those labs continue to operate in support of the analytic and research needs of states and districts around the country. And the new diagnostic measures that my colleagues from REL Mid-Atlantic will be describing in this session, which we developed in collaboration with states and school districts in the mid-Atlantic region, they represent the kind of creative thinking on policy-relevant measurement that, in my view, is going to be even more important during and after this pandemic concludes.

So, a brief summary of the agenda for the next 90 minutes. In a minute, I'm going to turn it over to Chris Boccanfuso, our project officer from the U.S. Department of Education, to talk a little bit about the REL. Then we'll spend the bulk of the time hearing about three different projects that REL Mid-Atlantic conducted to develop diagnostic measures. The first of these is on social emotional learning measures that we developed in collaboration with the D.C. Public Schools. The second makes use of kindergarten entry assessments to track the progress of the City of Philadelphia in promoting reading proficiency. And the third one also looks at the early elementary grades, using the growth from kindergarten entry to third grade to produce measures of early elementary school performance, which is a project we conducted with the Maryland State Department of Education.

Then, finally, we'll have some time at the end for discussion and questions. We have Joy Lesnick from the School District of Philadelphia as a discussant to start that conversation off, and then we'll look forward to hearing from your questions. So, with that, I will stop and turn it over to Chris Boccanfuso. Go ahead, Chris.

**[Chris Boccanfuso]:** All right, thanks, Brian. First and foremost, I want to echo Brian in thanking all of you who joined. It seems like the world is changing daily, if not hourly, due to the Corona virus. It has made all of our jobs and daily lives just that much more complicated, so we really do appreciate you for taking the time to tune in.

So, as Brian said, this presentation was developed in a pre-COVID-19 world. Clearly, there's going to be some issues with data collection for the current school year. That said, the fact remains that over the last two years, the amount of data available to states and districts has steadily increased. Available data is no longer limited just to state test scores, with a variety of data that can be used for informative purposes, such as data on early literacy, early numeracy, school climate, or student social emotional skills.

One of the big challenges with this additional data is becoming DRIP, or Data Rich, Information Poor, and that can be a challenge for any school district, even those with the most robust research and accountability department, to take advantage of the available data and to identify at-risk students and support formative improvement. This is really a space where the U.S. Department of Education's Regional Education Labs can be a catalyst for improvement across the classroom or in system levels. Regional Educational Labs, or RELs for short, work in partnership with schools, districts, and states in applying and developing actionable research and research evidence for specific problems of practice, with the ultimate goal of improving student and educator outcomes.

I quickly want to unpack the term “working with partnership” because the word “partnership with” can really mean a lot of different things. What I think makes the work of the RELs really special and really unique in the federal space is that it really is what I call a thought partnership, as opposed to being just a service provider. What I mean by thought partnership is that the RELs and the partners co-develop and execute projects through consistent, trusting, and mutually beneficial communication that values both research expertise and local context. This is done to ensure that work is actionable, relevant to stakeholders' context, and built to achieve stakeholders' goals.

This also means helping our partners answer the question of “What are our next steps?” once a project is finished. And so, I'm excited for REL Mid-Atlantic to share with you the results of three recently released studies that exemplify both the use of existing data to provide new actionable diagnostic measures and also true thought partnerships with three key stakeholders in our region. With that, I'm going to stop talking and I'm going to go ahead and turn it over to Kathleen and Tim to talk about the work REL Mid-Atlantic has done in partnership with DCPS.

**[Kathleen Feeney]:** Great. Thank you, Chris, and thank you everyone for joining us today virtually. My name is Kathleen Feeney, and I'm joined by my colleague, Tim Kautz. And we're going to discuss our partnership with the District of Columbia Public Schools in using their student-level, social and emotional learning data to develop what we called the Loved, Challenged, and Prepared Measure. This work was funded by REL Mid-Atlantic, and we'd like to thank our project officer, Chris Boccanfuso, for his support of the work, as well as the team at DCPS for their partnership, especially Sooyon Stiller, Emily Howell, Colin Taylor, Elizabeth Kim, and Brandee Tate. So, there we are.

I'll begin our talk today with a description of this project's purpose and how we partnered with the District. So, I first wanted to address two questions you might have. What are social and emotional learning, or SEL, competencies and why do we believe it's important to care about them?

SEL competencies can be generally designed as the broad set of competencies that are not adequately captured by achievement tests. Two that many people are familiar with are perseverance and self-efficacy. SEL competencies can also been known as character traits or non-cognitive skills, but we won't be drawing any sharp distinction between those terms today.

And while those characteristics have largely been overlooked, recent research shows that these competencies do rival academic achievement in predicting success for many critical life outcomes, such as educational attainment, health, employment, and earnings, and that these competencies appear to be malleable through the K-through-12 years and can be developed and improved through intervention.

So, essentially, the evidence shows that focusing on improving SEL competencies is a promising new avenue for improving the lives of children. And the team at DC Public Schools recognized this promise, and they wanted to act on it. So, inspired by this type of evidence, they highlighted SEL competencies as key outcomes for their current five-year strategic plan priority to educate the whole child. And this included the ambitious goal of wanting 100 percent of their students to feel loved, challenged, and prepared by the end of this five-year strategic cycle in 2022.

So, in order to track progress towards the strategic goals, DCPS wanted to know how to best use SEL data to identify student groups who need additional supports; track progress in students, schools, and the district as a whole; and evaluate SEL-related practices and programs. So, to that end, DCPS began administering a district-wide survey in spring 2018 to measure SEL competencies.

The survey was administered to DCPS students in grades 3 through 12, and covered both SEL competencies, such as perseverance and self-efficacy, as well as SEL supports, or in other words, the environmental factors that support SEL competencies. It included supports such as rigorous expectations or how much students feel teachers expect from them and encourage them. The survey asked students the extent to which they agreed to questions, such as if you fail on an important goal, how likely are you to try again, with a 5-point Likert scale of responses related to that question. For the first year of the administration, the response rate was 72 percent of students from the over 20,000 District students in eligible grades, and while complementary teacher and parent surveys were also administered as part of this effort, the data from those surveys were not used in the work we're focusing on today.

And while DCPS had collected this great data, there's really little formal guidance that currently exists on how districts can effectively use this type of data to inform their policies and practices, which brings us to the primary goal of this project and our partnerships with DCPS: to develop an index to measure the extent to which students feel loved, challenged, and prepared to align with their strategic goal.

So, DCPS partnered with Mathematica through the REL Mid-Atlantic under what's called a "coaching task," to explore the properties of their survey data. The coaching task emphasized two important goals. The first is a focus on capacity building, which meant that all analysis co-developed was shared with the DCPS team so that they could continue this work beyond the life of this project; and second was an emphasis on collaboration, meaning that the district team shaped decisions that we made throughout, especially as it pertained to stakeholder transparency. DCPS's priorities for the index really aligned well with the format of the coaching task, as the team wanted to create an index that would be both useful and usable by current and future district staff, as well as one that would be transparent so that they could give clear information to their stakeholders.

So, with these priorities in mind, the research questions that we addressed in creating this index were: Is the survey reliable and valid for DCPS students? What business rules should we use to construct an index that is transparent, while also capturing DCPS' constructs of loved, challenged, and prepared? And finally, how sensitive is the index to decisions about these established rules? So, at this time, I will turn it over to my colleague, Tim, to discuss our methods and findings.

**[Tim Kautz]:** Great. Thanks so much, Kathleen. And I'm going to talk through a four-step process that we followed to develop the loved, challenged, and prepared index, and when I say "we" here, I want to be really clear that I mean both the REL team and DCPS, because we developed this, really, through a close partnership, thought partnership with our DCPS colleagues in each of these steps.

So, the first step was to validate the survey for use among students in DCPS, and we felt this was really important because it allowed us to confirm that the survey measured what it was desired to measure and would be suitable for an index. I should say that Panorama has previously conducted some validation

analyses, but we thought it was important to conduct one specifically for this project for two reasons. First, DCPS included a few modifications to the previously validated surveys; and, second, DCPS' population of students might differ somewhat from those of Panorama's original validation analyses, and both of these factors could lead to different psychometric properties.

To validate the survey, we did a couple of things. We examined results from the confirmatory factor model, as well as some other types of specifics. But importantly, before conducting any of these analyses, we pre-specified some criteria for whether we would view the results of the analyses, and whether the survey would be viewed as reliable and valid. And if these criteria were not met, we would have considered some ways of adjusting the measures through liability.

Our analyses addressed two main questions. We first examined whether the items for each of the different topics of the survey consistently measured the same underlying construct, which is known as internal consistency, one form of reliability. So, for example, we looked at whether all of the items that were in to measure perseverance were related to other. And to assess this, we calculated Cronbach alpha, which is a standard measure of internal [consistency]. And for all of the different topics, we found that the Cronbach's alpha exceeded .7 for all of our topics, which met our prespecified criteria. And this is also pretty similar to what Panorama found in their original validation studies.

We also looked at the relationship between each item and each topic, and the underlying construct through a measure known as the "factor loading," which is basically an association between those two. And so, the higher the factor loading the more related an item is to the underlying construct. And we found that all of the factor loadings were sufficiently high to just say this item was the topic they were [inaudible].

Second, we examined whether the general grouping of items into different topics fit the data well overall. And to assess this, we looked at results from our confirmatory factor model. The basic idea here was to make a set of assumptions about the relationships between the items and the topics based on theory and see whether those assumptions matched the data by examining some specifics. And, in particular, we specified a model in which we assumed that each item that was designed to measure a particular topic only related to that topic and no other topics, and then we could test whether that grouping of items fit the data as well.

And to assess this, we looked at three different statistics and across these three commonly used statistics, we found that they met standards that have been used in the [inaudible], suggesting that the grouping of items into topics fit the actual data well. And so combined, these analyses gave us the confidence that the survey was measuring what it was designed to measure.

The second step was to determine which topics in the survey we would use to measure each of the three components of the LCP index, loved, challenged, and prepared. And we ultimately came to the mapping that's displayed in this table, which is based on a combination of theory and empirical evidence. The first, we captured whether a student felt loved with a topic called sense of belonging that summarizes sentiment to which students feel understood and accepted by others at their school. Second, we captured whether students felt challenges with the measure of rigorous expectations that Kathleen mentioned earlier, and this basically summarizes whether teachers have high expectations for students and whether they encourage them to keep trying if they failed. And finally, we captured whether students feel prepared using three different topics. These included perseverance, which relates to the ability to complete tasks and achieve goals; self-management, which relates to the ability to regulate behaviors and emotions; and self-efficacy, which relates to students' confidence in the classroom. And we focused on these three because they have all been linked to academic success or other long-term outcomes, suggesting that they capture some concept of preparedness.

The third step was to go from responses on each of the survey topics to a set of business rules that could be used to summarize whether students felt loved, challenged, and prepared. And one really important consideration when thinking about this was we wanted to create an index such that you could report separately each of the three components, so the percent of student that feel loved for example, as well as the percent that feel loved, challenged, and prepared in a way that's consistent. In addition, because these

would be presented to a broad set of DCPS stakeholders, we wanted to create as simple and transparent of an index as possible that would still maintain the necessary rigor.

So, to accomplish those goals, we settled on a set of business rules, which followed, of course, four key steps. So, for each student, we calculated the average response on a one to-five-point scale for each component in the loved, challenged, and prepared index. So, for example, one student could get a 3.5 on the loved component. Then, for each of those components, we determined whether that score exceeded a threshold, which we needed to represent a positive outcome for one of those three components that the threshold of 3.5 out of 5. And this allowed us to say whether an individual student, for example, felt loved. And then we defined the overall index for a given student on loved, challenged, and prepared to be whether a student felt loved, challenged, or prepared as defined by the step two. And finally, this allowed us to average across the students to determine the district-level loved, challenged, or prepared index, or the percentage of students who feel loved, challenged, and prepared.

The fourth step was to provide some evidence of specific decisions when actually calculating that score, and we explored a number of different issues. But today I'm going to talk about two of the key ones. The first decision was how to weight each of those three topics that entered that preparedness component, perseverance, self-management, and self-efficacy. And, for example, if evidence suggested that one of those components of preparedness was more important than the others, that could have received a higher weight when we were calculating student scores.

To explore this issue, we looked at two types of evidence. The first was how each of these topics related to data in DCPS' administrative records. So, for example, we looked at the correlation between each of these three topics and student outcomes, like achievement. And from that, we found that self-management and self-efficacy tended to be the most associated with student outcomes compared to perseverance.

However, we recognized that preparedness matters more across not just in schools but also for later outcomes, so we looked at the general literature on which of these three topics is most predictive of longer-term outcomes for students, and we found that perseverance was the best predictor. And this evidence suggested that each of these three components was important in different ways, and we ultimately determined to weight each of the components equally in the index.

The second decision was whether and how we should account for potential non-response bias. So, the goal of the index is to represent all students in the district. But not all students in the district respond to the survey, and if only certain types of students responded, then the overall index might not be representative. And so we did a few things to investigate this. First, we looked at whether the students who did respond differed systematically from those who didn't, and we could do this because we worked with DCPS to look at other administrative records that we had on all students. And we did find some evidence that different types of students were more likely to respond. And this caused us to take a second step, which was to develop a set of non-response weights, which allowed us to re-weight the data in the index to be representative of the population, of DCPS's population. And what we found when we did that is using those types of weights didn't really make a big difference in the overall index, and for the sake of transparency, we opted not to use those weights when calculating the final index.

And a kind of related issue was that some students might not take the survey seriously and exhibit what we called a low-effort response pattern. And we did a series of analyses, which I don't have time to go into, and we found pretty limited evidence that low response would make a difference to the index. So, those were the four steps that we used to generate this index, and I'll turn it back to Kathleen now to talk about some of the results.

**[Kathleen Feeney]:** Great. Thank you, Tim. So, in the end, I'm happy to report that we were successful in creating a valid Loved, Challenged, and Prepared index. So, the graphic shown here was produced by DCPS based on the findings from their first year of SEL data collection and was disseminated as part of their public year-one update towards their strategic plan goals.

You can see here that it shows the degree to which students feel loved, challenged, prepared, and also a combination of all three categories. So, as evidence of the success of the coaching model, the findings shown here from the LCP index are concise and are easy for stakeholders to interpret, and DCPS has since produced another round of findings based on their spring 2019 data.

So, to wrap up our presentation today, we also wanted to highlight some of the complementary work that we completed in addition to creating the LCP index, as well as where this work is going next. So, two other activities that we tackled as a part of the coaching task included exploring the properties of the complementary teacher surveys that I mentioned previously, as well as using natural language processing techniques to highlight themes in thousands of open-ended responses collected from parents.

And finally, as an extension of this work, we're partnering with DCPS again in a study that will explore how they can use their SEL data in other ways to track progress towards other 2022 strategic goals. Two main goals for the future work include studying how SEL competencies predict key educational outcomes, such as grade transition and graduation, as well as an investigation of the extent to which reports of SEL scales change for students over grades and how that might inform interpretation of SEL data.

So, I wanted to close by saying that we feel so fortunate to have been able to continue our partnership with DCPS, and we're excited to find what we learn from our next project together. So, this brings me to the conclusion of our presentation. Thanks again to everyone for joining us today. We look forward to your questions at the end of the panel, and with that, I'll turn it back over to Brian to introduce the next presentation.

**[Brian Gill]:** Thanks so much, Kathleen, and thanks, Tim. I think some of you have already started putting questions into the Q&A box. We're going to hold all the questions until after we get through all three presentations, just to make sure we're actually able to do that. But you can feel free, at any time, to submit a question for any of our presenters in the Q&A box in presentation. So, please, feel free to submit them, and we'll try to address as many as we can, turning some questions back to Tim and Kathleen when we get to the end.

Now, however, I'm going to turn it over to my colleagues, Jess Harding and Mariesa Herrmann, who are going to talk about some partnership work that we did with the School District of Philadelphia using kindergarten entry assessments. So, Jess and Mariesa, all yours.

**[Jess Harding]:** Great. Thank you, Brian. This is Jess Harding and, as Brian said, my colleague, Mariesa Herrmann, and I are going to talk about a research partnership with the School District of Philadelphia. We used a kindergarten entry assessment to track the city's progress in improving the number of students who can read in grade three. We'd like to thank our partners at the school district, Kristyn Stewart and Katherine Mosher, as well as our other colleagues at Mathematica, Elias Hanno and Christine Ross.

Kindergarten entry assessments, which are used to measure children's early skills, have the potential to be really important measures of children's outcomes because they're often the first large-scale data collection of young children's skills. In the past, they have been used to inform policy-makers' decisions about early learnings systems and to identify children's skills to guide teachers.

Here, we present another potential use for kindergarten entry assessment. The school district had a lot of efforts going on, with the goal of increasing the percentage of children who could read proficiently by fourth grade, and they wanted to track progress towards increasing the percentage of students who can read on grade level. But they would have to wait a long time to figure out if they were making progress if they had to rely on the grade-three assessment.

So, the district wanted to see if they could identify an indicator based on the first assessment given to students, called the "Pennsylvania Kindergarten Entry Inventory" in grade three, which we will call the "KEI." And so, they wanted to see if they could use an indicator, based on the KEI, to track the percentage of children entering kindergarten who are on track to be proficient in later years. This was a partnership where we worked with staff at the school districts to conduct the research, with them providing feedback on all

decisions. As Kathleen spoke about, we then provided coaching and technical support, as well as the statistical code to the districts, so they can run these analyses themselves in the future.

So, question one, we wondered -- we asked “What is the relationship between the KEI and the grade-three Pennsylvania System of School Assessment in English language arts?”, which we will call the PSSA. And the idea here is that there needs to be a relationship between these two assessments if we want to use the KEI to predict whether students are proficient in later reading.

For research question two, we asked “What threshold score on the KEI most accurately predicts proficiency on the PSSA for the cohort of students?” And, here, we tried to find a threshold that accurately predicts the proportion of students who are actually proficient in reading. As we mentioned earlier, if we can find an accurate threshold, then the district can know at the start of kindergarten that about, say, 40 percent of students are expected to be proficient instead of waiting until grade 3 to find out that 40 percent of students were proficient. So, this would allow them to track progress and early investment like early childhood education.

In some cases, it might be particularly important to identify students who were at risk of not being proficient in reading so that we could provide them resources to help them be able to read. So, we also looked for a KEI threshold that accurately predicts 90 percent of students who are at risk of not being proficient in reading. Finally, for research question three, the district also uses the AIMSweb assessment in reading in grades K through grade 3, so we look at how it relates to children's skills at kindergarten entry and their reading proficiency in grade 3.

We have four sources of data for children who entered kindergarten. We have the teacher-rated kindergarten entry, which are two validated dimensions I'll mention: emerging academic competencies, which measures early reading and math skills, and learning engagement competencies, which measures early behavioral skills. We also have the Pennsylvania System of School Assessment, the PSSA, which is a statewide assessment of reading in grade 3. We have information on student characteristics from enrollment records and, as I mentioned, we have the AIMSweb assessment, which is a reading assessment using K through grade 3.

For our sample of kindergarteners to be included in analysis, students had to have both the KEI and PSSA scores. And, ideally, we would have both of these scores for all students, and the closer we added that, the more likely it is that our sample will represent all kindergarteners. However, only 58 percent of kindergarteners had KEI scores. This was the first year of administration for the KEI in Philadelphia, so some teachers may not have administered the assessment. In addition, only 65 percent of those with KEI scores also had PSSA scores. Children may not have had those PSSA scores because they leave the district, enroll in charter schools, or are otherwise absent on the day of the assessment.

Because the district will use the threshold with all students who have KEI scores, we used weights to make the analysis sample, those with PSSA and KEI scores, more similar to all students who had KEI scores. And we did that weighting based on demographic characteristics and KEI scores. And we found that our weighting made the true sample similar to one another. Then, following the usual process for predictive analyses, we used one subset of data to get the threshold, and we used the other subset of data to test the threshold, and that helps us to check that the results aren't unique to the very specific sample that was used.

So, here, are the results for research question one. And we're predicting PSSA proficiency based on both of the kindergarten entry and [inaudible] dimensions. And what we find is that only emerging academic competencies predicts PSSA proficiency when both dimensions are included together. This is likely because emerging academic competencies includes literacy and math skills, whereas learning engagement competencies includes behavioral skills.

The average marginal effects column means that for every one-point increase in emerging academic competencies, the probability of proficiency increases, on average, 24 percent, and for every one-point

increase in learning engagement competencies, the probability of proficiency increases two percent, on average.

As you'll see, we'll compare how well each KEI dimension predicts proficiency, and how well an index of both dimensions together predicts proficiency. So, the coefficient column shows the coefficients that we used to construct the index of both KEIs I mentioned. So, that index weights that I mentioned according to their predictive power in predicting third-grade reading proficiency. Now I'm turn it over to Mariesa.

**[Mariesa Herrmann]:** Thank, Jess. So, to determine which measure to use, the emerging academic competencies dimension, the learning engagement competencies dimension, or the index in two dimensions that Jess just discussed, we calculated the percentage of students who would be correctly identified as proficient and not proficient on all possible thresholds on each measure. And so, this graph is showing those percentages for each of the three measures. And the green dot illustrates perfect predictions, which is where we would perfectly predict a hundred percent of the proficient students and a hundred percent of the not proficient students, while the grey dash line shows what would happen if we just randomly guessed proficiency of students; for example, based on a coin flip.

So, what you can see is, for all three measures, the lines curve towards that green prediction dot, which means that using these measures are better than just guessing, better than random chance. The emerging academic competencies dimension, which is the orange curve, and the index of the two dimensions, which is on the dotted black line, perform similarly well. The learning engagement competencies dimension performs worse. That's the one in the blue dash line, and that's consistent with its lower correlation with the PSSA.

So, since the emerging academic competencies dimension and the index of the two dimensions performed similarly, SDP selected the index of the two dimensions to use for the threshold because it uses data on more students. And so, after we select the index, we need to select the threshold. And if you focus on the curve, the black dash line, which is the index of the two dimensions, what the graph shows you is that as we try and increase the percentage of not proficient students who are correctly identified, we'll decrease the percentage of proficient students who are correctly identified. So, when we select this threshold, there's this trade off. Next slide.

So, because SDP wanted to measure the percentage of students who are on track to be proficient in grade 3, we wanted to select a threshold that would match the actual proficiency rate in grade 3 on the PSSA, which is about 37 percent, as closely as possible. And so what we did was we used one portion of our data -- that's the top row in the table -- to set a threshold of six on the one-to-ten index of the two dimensions that we constructed, and that's predicting a proficiency rate of about 37 percent.

So, to look at how well the threshold would work in data that weren't used to set it, we used the other portion of our data, which is the second row in this table, to test how well the threshold performed. And so, what that shows us is the threshold was similarly accurate in these data that weren't used to set the threshold. Next slide.

So, in addition to understanding how well the threshold performed overall on average, we wanted to understand how accurately the threshold predicted the proficiency of individual students. And so, this graph shows our one-to-ten index score on the X axis, and the vertical line shows our threshold, which is at six. So, the purple areas in this figure show the percentages of students who are actually proficient at each of the different KEI scores, and overall, the purple areas represent about 37 percent of students. The orange areas show the percentages of students who are actually not proficient.

And so, if you look above the threshold, what our threshold does it is predicts all students above it to be proficient. So, these predictions are correct for the students that are the darker purple area, but they're incorrect for the students in the light orange area. So, the threshold of six correctly predicts about 53 percent of proficient students, and that's the proficient students in the purple area of the figure.

So, if we look below the threshold, all the students are predicted to not be proficient. But, again, those predictions are only correct for some of the students, and that's the students in the dark orange area. So, among students who are not proficient, this threshold correctly predicts 73 percent of the not-proficient students.

So, even though the threshold is accurate on average, that's because the area above the threshold is the same as the area of all the proficient students, which are the two colored regions in purple, and the area is cancelled out. So, in other words, the light orange area that is above the threshold and the light purple area that's below the threshold are about equal. Next slide.

So, it might be important to correctly identify not-proficient students to help ensure that they receive the needed support, so we also looked at what threshold would correctly identify at least 90 percent of not-proficient students, which is a criteria that's suggested by prior research. So, this graph shows what happens if you increase the threshold to seven, which correctly identifies at least 91 percent of not-proficient students. And if you shift the threshold right, it increases the percent of not-proficient students who you correctly identify, so that large orange region expands. But this also decreases the percentage of proficient students, which means that you have a smaller dark purple area. And, in fact, this threshold only correctly identifies 29 percent of proficient students.

So, in addition to the KEI, SDP also administers the AIMSweb reading assessments kindergarten through grade 3, though they did not administer it in the fall of the kindergarten for the cohort we examined. So, to understand how the AIMSweb assessment might compare to the KEI, we calculated the correlations between the AIMSweb in spring of each year and proficiency on the PSSA. So, this graph shows you the relationships between each assessment and the grade 3 PSSA score.

At the top, we have the correlations for the two KEI dimensions that were used in the index, and at the bottom, we have AIMSweb scores in spring of kindergarten through grade 3. And so, you can see the strength of the associations are higher for AIMSweb in the spring of kindergarten relative to KEI in the fall, and the associations are also stronger for AIMSweb scores that are closer to grade 3. And so, this suggests that AIMSweb scores could be more accurate measures for tracking students' reading proficiency, particularly as you're getting closer to grade 3.

So, we wanted to highlight some limitations of the analysis. So, first, some of the students with KEI scores did not have PSSA scores. So, for example, because they were absent on a test day or they moved out of the district. So, we adjusted for this in our analysis using weight, but it's possible that students without PSSA scores could have had worse outcomes than students that otherwise looked similar to them in kindergarten.

So, second, we also set the threshold using one cohort of students, those that entered kindergarten in the '14/'15 school year and we predicted their proficiency in grade 3. It's possible that the threshold might not have the same accuracy for other cohorts if the assessments change, so, for example, SDP was providing more training to kindergarten teachers on administering the KEI. And finally, we looked at two validated dimensions of the KEI, the emerging academic competencies and learning engagement competencies dimension. Other KEI items or other assessments like the AIMSweb could have strong relationships with PSSA proficiency.

So, in terms of implications and next steps, first of all, the study found that scores on Pennsylvania's kindergarten entry inventory, particularly on the emerging competencies dimension, is related to proficiency on PSSA in third grade, so this suggests it's possible to set a threshold that accurately predicts the cohort's proficiency rate. And we also learned that a different threshold that predicts a higher percent of students that are not proficient might be more appropriate for identifying students who need support in reading.

Also, as we noted, these findings are based on one cohort of kindergarteners, and because this was a technical assistance project, we provided SDP with a code, and so they plan to reassess the accuracy of the threshold on the KEI using data on additional cohorts. And we also noted that we found some stronger associations between the PSSA and the AIMSweb assessments that were taken at the spring of

kindergarten, so they'll also be able to look at AIMSweb assessments from fall of kindergarten and consider whether it could make sense to set a threshold on those assessments.

And finally, for states and districts, the findings suggest that kindergarten entry assessments could have additional uses beyond sort of tailoring instruction for individual students. So, for example, you could use the data to track the progress of cohorts of kindergarteners towards goals or target supports of schools where students may be at risk of not reading proficiently.

**[Brian Gill]:** Thanks very much, Mariesa. Thanks, Jess. So, we'll turn now to the third and final of our presentations on the development of new measures here. And, again, as before, please feel free to submit your questions for any of our presenters via the Q&A box. We will get to them shortly, when we get to the end of this.

This last presentation, like the one you just heard, is also related to analyzing kindergarten entry assessments and that relationship to third grade statement scores. Although, in this case it was statewide data from Maryland rather than citywide data, as in Philadelphia, and the intention was a little bit different, trying to assess the overall preparation of a cohort of students from an entire city. In this case, the question was whether it's possible to produce valid and reliable growth measures for individual elementary schools for achievement growth for students from grades K through 3. So, my colleague, Dallas Dotter, is going to talk to us about that. All yours, Dallas.

**[Dallas Dotter]:** Great. Thanks Brian. So, I'll be presenting the results from a study that we performed as part of a partnership between the Maryland State Department of Education and the REL Mid-Atlantic. And, really, this was a study that came out of Maryland's desire to understand whether or not they could measure student growth in the early elementary grade levels. Academic growth is one of many measures against which schools are evaluated for accountability purposes, and in Maryland, it accounts for a quarter of a school's accountability points. But, as is the case in most states, students are tested with a PARCC assessment in Maryland in grades 4 to 8, so measuring student growth is only possible in the later elementary grade levels.

However, there was a kindergarten readiness assessment that Maryland implemented in the fall of 2014, and that was a potential to provide a baseline measure from which growth could be measured to grade 3. So, this partnership really looked at whether or not it was going to be feasible to use that assessment to measure growth to grade 3 and look at the properties of that growth measure.

So, the study samples looked at the first cohort of kindergarteners to take the KRA, Kindergarten Readiness Assessment, and this was in the fall of 2014. The sample for the study included all the students who took that exam who also had grade 3 PARCC scores in math or reading. This ended up being 86 percent of all students who took the KRA in 2014. This included all schools with kindergarteners and third graders, 914 schools in total, and we estimated student growth percentiles using these scores, which is also the measure that is used for student growth in the later grade levels in Maryland.

To calculate student growth for this study, we aggregated them at the school level using the mean of student-level growth percentiles, and because there's a longer time period between kindergarten and third grade than in the year-to-year student growth percentiles calculated in grades 4 to 8, we accounted for students attending different schools in that timeframe using weights proportional to the number of years they attended each school. The estimates that we produced for student growth percentiles accounted for measurement error in the kindergarten readiness assessment.

The study looked at four main research questions. The first was whether an overall KRA score that's reported would make the best baseline measure for the two-growth percentiles or whether some combination of the four sub-scores that are reported for the KRA might perform better as a baseline measure. We also looked at whether the K-3 growth estimates were valid and precise relative to the student growth percentiles that are produced for later grades. What we mean by "valid" here is that what is being measured by the two assessments are similar and measuring growth from one to the other is providing something that's meaningful. We also looked at the precision of these estimates. In other word, holding a

school's performance constant over time, how would these vary year to year as different students attend school.

The third question we looked at was how the precision of these estimates was affected by the size of a school, and the last question was how administering the KRA to a smaller sub-sample of students would affect precision of their assessments. And this last one was really motivated by the fact that the KRA was administered to a sub-sample, roughly a third of kindergarteners in some of the later years.

So, for the first question, what we found is that the KRA overall score as a baseline measure performed about as well as the other scores we looked at. So, we looked at the overall score and we looked at using the same subject domain scores, that was the math domain score for the math SDP model and reading for reading. We also looked at some weighted combinations of those scores. One was a weighted combination of the math and reading sub scores. The other was a weighted combination of all four of the domain scores, and those weights were produced using a regression cross-validated, essentially, or intuitively, producing weights that would maximize the predictive power of a student's KRA score for their grade 3 score. So, given that there was little difference, looking at the correlation of this table on the right, between the different measures of the KRA as a baseline measure, we proceeded with the study to use just the overall scale score for transparency and ease of interpretation.

Next, we look at the correlations between KRA and grade 3 scores, really, to look at whether these two assessments are strongly related within student performance. We found that the correlations between these within students was substantially lower than the correlations between grade 3 and grade 4 performance on the PARCC. For the KRA in grade 3, PARCC math and reading correlations were the same at .53. The correlations were quite higher for students grade 3 and grade 4 PARCCs, between .77 and .87. Even though these are lower, these are sort of in the range that we might expect from the literature. Looking at the literature, correlating performance on KRA assessments with later assessment in grade 3 ranges anywhere from between .1 to about .7, so these correlations of .3 are towards the higher end of this range.

So, to look at the precision of these K-to-3 SGPs, we used a bootstrap procedure, where we resampled students within the same school with replacement and produced thousands of these samples with different random draws of the students that had been taught by schools. Each time we re-estimated student-to-student growth percentiles statewide, and then reconstructed a school's mean SGP. So, repeating this, we're able to build a distribution of SGPs for each school, and we constructed 95 percent confidence intervals using these distributions. So, what we found was that the width of these confidence intervals was about 12 percentile points in math and 13 in reading, and this is very similar to the confidence intervals that we find for grade 3 to 4 SGP when we estimate those using the same procedure.

Next, we looked at correlations between K-to- 3 SGPs and schools SGPs for grades 3 to 4, math and reading. And the purpose of this exercise was to see whether measuring growth from kindergarten grade 3 provided roughly the same information as measuring growth for schools in grades 4 and higher. What we found is that these were moderate correlations for schools, so a high correlation here, close to one, would mean that schools performing very well or producing large amounts of growth, the students would do the same in the higher grade levels as in the lower grades, between kindergarten and grade 3. These moderate correlations suggest that measuring growth between kindergarten and grade 3 is providing some new information that might not be captured of students' growth when one just relies on looking at SGPs in grades 4 and higher.

Looking at the relationship between the size of schools and the precision of these K-to-3 SGP estimates, we see that schools that are particularly small can have quite wide confidence intervals. From this figure, these would be the schools to the left of the graph, and there's a wide range of confidence intervals for these schools. But after roughly 50 students within a school, these confidence intervals are always within 20, and after a hundred, they seem to converge around a confidence interval around 10 percentile points.

Finally, we looked at how precision would be affected by a partial KRA administration, which happened in later years. For our year, the fall of 2014, the first cohort that was administered to the KRA, there was a sense in the administration where every student took the exam in the fall. In later years, this varied between

34 to 39 percent. And so, going back to the bootstrap exercise we did, free samples students within schools, we simulated this by sampling not the original size of the school but a proportion equal to these percentages. What we found is that these partial administrations would roughly double the width of the confidence intervals. So, looking at this distributional graph on the right, we see that most of the schools have a confidence interval of 15 percentile points or fewer under the [inaudible] administration in red, whereas the majority of schools have a confidence interval 15-point swing, this partial administration would take place.

So, to wrap up, I'll summarize some of the limitations of the study. First of all, the findings here only reflect students in the analysis, obviously. As I mentioned earlier, of the students that took the KRA in 2014, 14 percent of them did not have a grade 3 score. Most of those had exited public schools by that time, and, secondly, there are no KRA scores for grade-3 students who had entered Maryland public schools after kindergarten.

Secondly, the reliability of these estimates could not be studied. The data we had included the first cohort that took the KRA as the later year to take the grade 3 part, so future analyses could be conducted to look at how these within school SGP estimates might change over time.

Third, while we accounted for students' moving between schools, between the grades of kindergarten and grade 3, we didn't observe within school year movement, only between school years. So, our weight that account for movement could be improved if the data were a little bit more fine grained and could follow students within the school year.

Fourth, we use the PARCC exam as a grade three score here, but that's going to change in Maryland. So, all these results here are pretty informative of what one can expect from using the KRA and these grade-3 PARCC scores, some of these results may change as the exam moves to another assessment.

So, these findings were pretty interesting. Maryland was definitely interested to see the relationship between the KRA and the PARCC, and the statistical properties of producing growth measure out of these. This growth measure shows that it can provide new information about growth that isn't captured in some of these later grade SGPs and grades 4 through 8. This may be a model for other states to look at. Perhaps looking at other kindergarten readiness assessments or kindergarten assessments or assessments done in grade 1, 2, or 3.

But, as we saw before, I think there are a few things that one should be cautious about when you see these measures. We did see that the correlations were quite lower between the kindergarten assessment and grade 3 PARCC scores when you're comparing that to the correlations of students in grade 3 and grade 4 scores. And also, the precision was quite low for some of these small schools, which suggests that one might consider using some sort of threshold on school size for attributing these measures, or some other way to account for low precision, like empirical data shrinkage. So, that concludes the results for our study, and I'll turn it back over to Brian.

**[Brian Gill]:** All right. Thanks very much, Dallas. So, we want to get to the discussion now and address any questions you might have. Again, feel free to put your questions in the Q&A box, questions that you might have for the presenters or any of our guests here. And to kick off our discussion, I'm going to turn it over to Joy Lesnick from the School District of Philadelphia. Go ahead, Joy.

**[Joy Lesnick]:** Thanks Brian. Good afternoon. Thanks everyone for joining. Thanks to the presenters for sharing your projects. My name is Joy Lesnick. I'm deputy chief of Research, Evaluation, and Academic Partnerships at the School District of Philadelphia. I've been in that role for three years, and prior to that, oversaw the REL program at IES, so I'm grateful to have different perspectives on these projects to inform some of my brief comments here as “discussant”, which I’ll put in quotes, for what was originally scheduled to be a SREE presentation, so I'm grateful for this opportunity.

First, I want to talk about how these three projects are really great examples of the value of working in research practice partnerships to use existing data to develop new diagnostic measures. So, first, Tim and

Kathleen talked about their work with D.C. Public Schools to use student-level social and emotional learning data to develop a loved, challenged, and prepared measure from existing survey data. And working closely with their DCPS partners, they used both existing survey data and the literature to work through some important decisions about the components of that loved, challenged, and prepared overall index, and the separate indices too, being loved, challenged, and prepared to help the district measure progress.

They also paid really close attention to non-response. The survey response rate was really high, 72 percent, but also paying attention to students who were not responding, and the implications for decision-makers in using that index. So, I think the answer to their main question, can the district social/emotional learning student survey be used to measure whether students feel loved, challenged, and prepared is yes, I think so.

We then heard another project, the REL Mid-Atlantic with my district, the School District of Philadelphia. Jess and Mariesa talked about a project that used existing kindergarten entry inventory, or what we call KEI assessment data, as an early warning indicator of sorts for third-grade reading proficiency, which is the city-wide shared goal. This was an interesting one from my perspective, of course. Our city partners, as well as our state Department of Education in Pennsylvania, who was responsible for the content of the KEI, we're especially interested in this analysis. Our third-grade reading proficiency rates are quite low in Philadelphia, around 37 percent, as the presenters discussed, annually, so it's often more challenging for teachers to provide grade-level instruction to large groups of students who are below grade level than to identify who is below grade level. But, nonetheless, exploring whether those kindergarten entry inventory domains could be a leading indicator of third grade proficiency, and possibly a tool for monitoring efforts to improve early or free kindergarten interventions and supports is valuable, especially before they enter the K-12 space.

We have a lot of community-based efforts to support students and families before students enter kindergarten, and this measure really could be used to monitor the progress of those efforts. Internally, I don't think we'll use the KEI itself as an early warning indicator of third grade performance. We have other assessments that fill that role. But it is the first measure of school readiness that students complete, and we're thinking about how to use it as a measure of preschool support or as a lead indicator for third grade. We have also used the threshold approach in other projects totally unrelated to the KEI, so that has been very valuable as well.

So, to keep the same format and the question I pose on the slides, can the kindergarten entry inventory be used as an early warning indicator of reading success in 3rd grade, I think the answer to that is yes, and, more importantly, it's probably a good measure of the leading indicator of early learning reports in the city before students get to kindergarten.

Finally, we heard from Dallas Dotter, who talked about a project with the Maryland State Department of Education to develop a measure of school performance for grades K to 3. And similar to the study in Philadelphia, they looked at the relationship between kindergarten readiness, the kindergarten readiness assessment to KRA, slightly different acronym, and third grade performance to determine if they could construct a measure of the school's contribution to K-3 growth.

And they found that the K-3 growth measure was different than similar measures for higher grades, specifically the third to fourth grade measure, so it did provide new measures. It did provide new information, but with a lot of cautions. It was less valid than the growth measure for grades 3 to 4, less precise for smaller schools, and less precise when the KRA wasn't administered completely to all kindergarteners.

So, to keep the format and answer the question I posed on the slide, is using the kindergarten readiness assessment a valid way to measure school contribution to K-3 growth, I think the answer here is not quite, but I welcome the authors and presenters all to disagree about my simplified questions and answers here. That said, whether the answer is yes for the first one, yes, but for the second one, or not quite for the third one, that answer is all valuable information. And using the existing data that we have in districts and states to answer those questions is an added bonus as well from my perspective.

The projects presented here also used some fairly sophisticated analyses. There are sophisticated analysts in states and districts, certainly, but it's also extremely useful when the research partner understands the practitioner context enough to help translate those sophisticated research findings into decisions. For example, the threshold graph that Jessica showed in the School District of Philadelphia KEI project was especially informative in our partnership when we talk through the results. Jess said the findings don't need to be as simple as a yes, no, or not quite answers, again, I mentioned, so that's helpful too. But clearly describing the nuance and considerations within time and attention limitations is hard to do, and I think these projects are really good examples of doing just that.

My final comments are about the challenges of being both proactive and reactive and asking and answering research questions. This is a topic I thought a lot about working in a district that has lots of questions about our own context and is committed to using research to inform decisions. We also expect or maybe perhaps hope that research has already been done on whatever topic comes up at that moment that we care about. These are all real examples: the outcomes of students when high and low performing schools are co-located; the impacts of hiring private staff for lunchtime transitions on disciplinary referrals; whether money is effectively spent by eliminating cross-breed split classes like combining second and third grade classes; effective ways to keep students safe without using metal detectors; and most recently, just two weeks ago, what the research says about teaching students remotely during a global pandemic. The research on that is quite thin.

In RPPs, we often place high value on questions that already exist so we're not working to create a new demand for the findings, and that's really what I like about applied research, and working inside a school district too. People typically want to know the answers to the questions they ask, and fast. But there's also an underlying demand for predicting questions that will come up in the future, and the answers we'll need. And using existing data to try to answer some of those not yet but soon to be important questions, such an important phase for researchers, for the RELs, for the RPPs, and beyond, especially when the [district] staffer is consumed with answering a pressing question of the moment.

So, I think we have some questions already in the chat box. If you have some, you can add them in there. I want to get us started perhaps and go back to the authors, and, certainly, if you disagree with my yes, yes but, and not quite answers, feel free to do so. And my question is really to talk about how you balance the sophisticated measure development you are doing with the need to make meaning that your practitioner partners were likely struggling with along the way. I will turn it back to Brian who, I think, will moderate this, and maybe some presenters will jump in.

**[Brian Gill]:** Okay. Thank you so much, Joy. I'm going to say, first, that I'm very glad that you find all of these as good examples of the value of working in research practice partnerships to use the existing data to develop new diagnostic measures, because, in my view, they also are good examples of how the Regional Labs can provide critical research and analysis for states and districts in developing systems to support continuous improvement. Because if states and districts are going to promote continuous improvement in their practices and, ultimately, improvements in teaching and learning, they not only need to know what works, but they also need good measures, what all these projects are about, and they need good measures for a couple of important reasons. One, so they can identify needs of students, educators, and schools; and two, so they can monitor progress and assess the effectiveness of interventions. And these are the kinds of measures that can start to get at those issues.

But let me ask any of our presenters, who would like to take Joy's first question about balancing this translation challenge? Tim, Kathleen, Jess, Mariesa, Dallas, who wants to tackle that? That could go to any of you, so do we have any volunteers?

**[Jess Harding]:** This is Jess.

**[Brian Gil]:** Go ahead, Jess.

**[Jess Harding]:** This is Jess, and I can take a first pass. I think, as Joy said, we were lucky in the school district, where our partners had a lot of sophisticated knowledge of research methods and how to conduct

these analyses as well. But we did try to ball balance the different presentations that we did according to the audience, so some of those being research folks, others being partners, to really try to make sure that the meaning and most relevant points were relevant for each presentation that we were doing.

**[Brian Gill]:** Great. Thank you. I'm going to ask Dallas to respond to the sort of input that we just heard from Joy, which is just ask, do you agree with Joy's characterization of the answer to this question as, well, maybe the findings show that kindergarten readiness assessment in Maryland doesn't allow a valid measure of schools' contribution to growth in grades K through 3?

**[Dallas Dotter]:** Yeah. No, I think it's a good question, and a reasonable assessment. There are certainly some things to be cautious about in a measure that has that many years between assessments and uses different assessments. The correlations we saw between the assessments were not super strong, and we saw that the precision, even though it's similar to precision for the SGPs used in later years, it can be high enough that sort of wonder where in the distribution of schools does a particular school actually land. I think, as with any measure of growth for accountability, there should be abundance of caution used in sort of how it is interpreted, and, also, making sure that it's not the only measure that's going into the evaluation of a school.

**[Brian Gill]:** Okay. Thanks. I guess I should say that I have, I suppose, a somewhat more optimistic interpretation of this, but maybe that is, in part, because the view of these measures depends, in part, on what other kinds of information is available. As you've just suggested, Dallas, it's useful to have multiple kinds of information. And what's interesting to me about the K-3 context is that there are no states anywhere in the country that actually know how much growth their schools are producing in those grades. You know, pretty nearly every state starts statewide testing at the end of third grade.

Kindergarten readiness assessments are becoming more common though, and so, in some sense, there is an opportunity here to make the learning in the early grades transparent in a way they haven't been before, which strikes me as, potentially, useful, you know, even realizing that there's some limitations of these measures -- that I read these finding as suggesting that they do provide some useful information, and that would be better than the nothing that's out there at the moment. But this is definitely something that different folks could read differently.

I want to turn to the questions that are coming in from all of you. The first one, very straight forward question for Kathleen, which is, “On the loved, challenged, and prepared measure”, someone asks, “what are the ages and grade levels of the students who are being measured for this?”

**[Kathleen Feeney]:** That's a great question. So, the SEL survey data is collected between grades 3 and 12. I should note that there are two slightly varying versions of the survey questions, one for grade 3 through 5 and the second for grade 6 through 12 to account for reading level.

**[Brian Gill]:** Thanks. Thanks much, Kathleen. Okay. Another question here that is relevant to both the Philadelphia study and the Maryland study, and so, Jess, Mariesa, or Dallas could address this. But someone asks -- I think this is a great question, particularly for people who are more familiar with assessments at higher grade levels. “How are these kindergarten entry assessments actually administered? Are these five-year-olds taking bubble tests?” Who wants to volunteer for that one?

**[Mariesa Herrmann]:** So, this is Mariesa. I can say something about this, and then if others, Jess or whoever wants to join in. So, my understanding is, in Philadelphia, the teachers administer the kindergarten inventory by observing kids doing a variety of classroom activities during the first 45 days in school. So they would, like, then rate them on a series of indicators like how well are they counting or, you know, their behavioral regulation, and they would give them a score on a 1-to-4 scale, like are they not yet evident, are they emerging, are they evident, or do they exceed what they're supposed to be doing. And then for our study, we're looking at these two validated dimensions, and so the dimensions are just the average ratings across the indicators that map to those dimensions.

**[Brian Gill]:** Thanks. Thanks, Mariesa. So anyway, for those of you who are wondering, quite different from a bubble test, and a lot more demanding, actually, in terms of the administration.

Another question related to kindergarten readiness, someone asks whether there are any links to data coming from preschools, recognizing that there are some preschools that do some similar kinds of assessments of their own, and wonders if K-12 schools ever have access to that information when kids enter kindergarten. So, let me just ask Joy, in Philadelphia, if the School District of Philadelphia has access to any sort of information that actually comes from preschools?

**[Joy Lesnick]:** We do have a connection between preschool teachers and kindergarten for places where there is a kind of clear feeder pattern. We have a variety of different ways that children attend preschool in Philadelphia, so those data systems tend to be separate. But we are working on getting those together in a more systematized way. One challenge is that students have a different ID in preschool than when they are K-12, so that's a data challenge to figure out, but it is something that we're working on to have that more clean transferred information from pre-K for students who attend to kindergarten.

**[Brian Gill]:** Okay. Thanks a lot. Okay. Now, we've got some questions on the Philadelphia research. I think at least the first of these probably should go to Mariesa. So, someone asked for a little more explanation about the tradeoffs between different thresholds in the assessments, and specifically if you wanted to use a threshold that predicted proficiency and got 90 percent of them, would that be a good idea, and what would be the downside trying to do it that way?

**[Mariesa Herrmann]:** I guess what I'm wondering is the question, what's the use of the threshold? So, we're kind of looking at two uses of threshold. One was to try to accurately predict the percentages, the cohort of kindergarteners who would be likely to be proficient in reading in grade 3. And for that, you know, we suggested using a threshold that didn't, in fact, predict 90 percent of proficient students, because in order to accurately predict it, you have to set the cohort rate. You have to set the threshold differently.

I'm wondering if this question is about the threshold that predicted 90 percent of the not-proficient students, which was the one that we looked at for the second research question. You know, would you want to predict 90 percent of the students who were not proficient or at risk of not being proficient in grade 3? The downside of that is that, you know, you may have limited resources as a district to, you know, provide support to students who are not going to be proficient. So, if you over-identify kids that at risk, then you may not be able to provide support to all the students. I think Joy kind of mentioned that, you know, it can be difficult to try to support all of the students in, like, a group setting, and so, in that case, you might want to, you know, be more accurate about trying to target support to the students who need it the most.

**[Brian Gill]:** Great. Thank you. Mariesa. Yeah, so this is, I think this study provides a nice illustration of how, you know, there's not a single threshold that is optimal for every purpose. And, in fact, given that you don't have a perfect assessment, you never have a perfect assessment, the threshold you want to choose really depends on what you're planning to do with the information and, indeed, there are reasons to think you might have different thresholds for different purposes.

Let's see, two follow-up questions from the same person. One is, do you have any way to guess -- I think it would only be guessing, given the limitations of the data -- how much cohort effects might matter, you know, how well this might predict for different cohorts? And then if you had more data across years and within years, would you recommend a different analytic model like a multi-level logit model?

**[Mariesa Herrmann]:** Okay, so I guess, Jess, did you want to take any of those?

**[Jess Harding]:** No, you go ahead.

**[Mariesa Herrmann]:** Okay. Sorry. So, in terms of the cohort affects, I mean, that's a great question. We're not really sure to what extent that there are cohort effects and how much year-to-year variability that there would be. And one of the goals of this project was to provide technical support to SDP so they could sort of

reassess the accuracy of the threshold when they have additional year of data. So, it would be interesting to see what they found on that.

With respect to the second question, if we had more data, would we use a different model? Well, for this study, we were interested in just setting a threshold based on these two dimensions, and so we did use a logit model. But we didn't do anything, like, try to nest things within schools or classrooms or something like that. And I think whether you did that would sort of depend on what you were trying to do, if you were trying to control for some sort of school effects or classroom effects. But we just wanted a threshold that one could apply across the entire district.

**[Brian Gill]:** Okay. Thanks, Mariesa. We have just a few minutes left for more questions. If you've got any, feel free to write them in. I think we have at least a couple still waiting. Let's see, one for Tim. We have a question about the natural language processing you did. Can you tell us a little bit about that analysis and what you learned about social/emotional learning through the examination of the open-ended responses?

**[Tim Kautz]:** Sure. So, the natural language processing part of this project was tied to the parent survey. So, as we mentioned earlier, there was the student, teacher, and parent survey. And parents had an opportunity to provide feedback for the district through an open-ended response, so they could comment on various topics. The challenge is that those kinds of -- because there are so many responses, DCPS was facing trouble trying to organize those responses in a way that they could make sense of it. And to do that, we applied a natural language processing algorithm, which did two things. One, it rated whether those responses tended to be positive or negative for each one so DCPS could easily tabulate the number of positive and negative responses, and the second is that it categorized the responses into different topics.

So, you know, one topic that came up was school climate. And so, by applying these algorithms, DCPS could then basically describe whether parents felt positively or negatively about a whole range of different topics in a way that was easy to summarize and didn't require a lot of people to read each open-ended response. Kathleen, is there anything you'd add to that?

**[Kathleen Feeney]:** No. I think that's a great summary. And as you say, I think, if my memory serves me, I believe there are over 3,000 individual answers that came in, so, you know, when you're thinking about all of the things that a district has already got on their plate, you know, having someone read each of those responses is just not feasible. So, I think this is a really interesting tool to get to use to see what kind of feedback the parents did have with the district.

**[Brian Gill]:** Great. Thank you, Kathleen, and thank you, Tim. So, we are about at time now, and I just want to put up the contact information for myself and all of our presenters here, so if anybody has follow-up questions, you can feel free to get in touch with us. And I want to thank you all again for joining us, in this really crazy time frankly. I hope you found this interesting work, and we look forward to connecting with you in other ways. Thanks very much everybody.

[End of webinar]