

Sanford

The teacher-to-teacher initiative was created by the U.S. Department of Education to provide the latest strategies and research on educational practices that work inside a classroom. Howdy, my name is Sandy Sanford and I'm from Riverside County, California. What I want to talk to you about today is a journey that we have been on for the past four years. Which is more important for you in the classroom, Norm-Referenced Results or Criterion-Referenced Results? If you're really a superb fourth grade teacher, you're not only in touch with the state content standards for fourth grade, but you are for third and fifth grade as well. This series features teachers from across the country presenting techniques that can be used with students of all ages. It's just one way the Department of Education is helping teachers get the support they need so no child is left behind.

Howdy, my name is Sandy Sanford and I'm from Riverside County, California. Now, what I want to talk to you about today is a journey that we have been on for the past four years to develop ways to excite teachers with data in order to support the goals of our state assessment accountability system and the goals of No Child Left Behind. The way it will be applicable to you is because I believe that the lessons we've learned and the techniques we've started piloting will be valuable in each one of your schools, your districts, your states, just as it was for us. Maybe what we've learned can save you a lot of time. Just to give you an example, the terms that are up here, kind of outline the situation as I'd like to describe it. In 1998, in California, I'm going to use my state for an example, but I think it applies equally everywhere. We started our standards-based education program. We invented the standards for the state as a whole. That was the first point in which we had a statewide articulated curriculum. Right after that, we went into a high stakes testing program to assess the results of that, and that has evolved and picked up energy with the advent of No Child Left Behind in 2001 and has evolved to where we are today. Part of the problem was that the teachers, during the course of the year, wanted some method to gauge the progress of students, according to the content standards with respect to where they are going so they can tell how they were progressing toward the high stakes testing that was going to occur at the end of the year and more importantly, much more importantly, to determine how they were doing at increasing student achievement. Now, I'm the perfect guy to give this presentation, I really am. I'm probably one of the most boring people in the world. I started in this assessment and teaching game in 1970 and I'm still in it, and I love it. I love data. I love assessment. I love measurement. I'm not so obsessed with it that I think it's the beat all to end all. I love my time in the classroom. I would not have graduated from high school if it hadn't been for the fine arts programs and the performing arts programs. I'm well aware that the curriculum and what we need to do to educate our students is much more than just what high stakes testing measures. But for the purpose of this presentation, I'm your perfect guy because I really love this stuff. When you're in bed at night, reading a novel, I'm reading a monograph on multi-colinearitic problems to do with multiple measures. I've never had a date with the same girl twice. (laughter) Anyway, my point being is that I am the perfect guy for this. I really love it. It's what my whole life has been about. I do it seven days a week, and so I'd like to share what I've learned with you. If you'll look down this list here, you'll notice things like data driven decisions and data to information, these are terms that are right off the slide that Dave Butler used earlier today when he was talking about No Child Left Behind and what some of the focuses of that are. The last point I'd like to make on this particular slide is metacognition. One

of our goals in developing better ways to use data in the classroom was to give metacognitive tools to the teachers, both from a teaching and analysis standpoint and from a learning standpoint, to put the students in the role of assessing their own metacognition. Now the outcomes for this particular presentation are here. We're going to first talk about data that informs and data that judges and why both are important. One critical change that's happened in the assessment game all the way across the country, with the advent of content standards in each of the states, is a movement from Norm-Referenced Tests to Criterion Referenced Tests as a primary way of evaluating and that's important to understand. I want to talk about what I consider the difference between testing and measurement and why that's important. We're going to talk about the characteristics of Teacher-Friendly Data. During the last four years of our research, much of the criticism that we have been addressing from the teachers, has been based on data not being friendly or usable to them, to us. And then, to describe something that we came up with called wrong answer analysis, and then the application of wrong answer analysis in a process that we call "WHAT", Why and How Analysis for Teaching. First thing I want to talk to you about is Data that Judges and Data that Informs. The metaphor I want to use is a basketball team. Now I put it in the form of sports metaphor because you too can use this metaphor when you're talking to your parents and to your peers. Picture a basketball team and this basketball team wants to get to the playoffs. There are various stakeholders involved with this basketball team, players, fans, the owner and the coach. Now this is an analogy and metaphorically the coach is the teacher, so keep your eye on the coach. Now as a data person, as a statistician if you will, I can describe this basketball team in several ways. One way is to describe it with something I'll call Data Set "A". Data Set "A" is exclusively the won/lost record. That's all it is, the won/lost record. Data Set "B", another way I can describe this team is the myriad of individual and team statistics, the free throw percentages, the penalties, the turnovers, both for the team and the individuals as a whole. Now, the players, in the heat of the game, when they're trying to get to the playoffs, this is not a rhetorical question, what data set are they most concerned with? Data Set "A", I mean they want to win the game; they want to win the game. Okay, how about the fans? "A" Okay, any doubt about the owner? Do I even need to ask that question? Now the coach, the coach is in kind of an unusual situation, when that coach, that teach if you will, is working between the games to make that team better than it was the previous game, to be the best team possible for the next game, which data set is most important? "B". In fact I think you could make a case for Data Set "A" is almost worthless to the coach in this situation. You see what I mean? The coach needs to know everything about that team he or she possibly can, in order to mold that team into a better team and to function together better than they were before. This is very comparable to all the records you keep in your classroom all the stuff on the individual students. The big difference is the coach can put this in the paper and you can't advertise it. Now, I call Data Set "A" data that judges, sometimes it's referred to as political data. And Data Set "B" is data that informs, sometimes referred to as educational data. Now an important part about this is just because, to the coach, Data Set "A" is not useful, for that particular mission, does that mean it's not important? It's extremely important. When that coach is negotiating his or her contract for the next year, which data set is most important? Data Set "A" Conversely, when each one of those players, as individuals, are negotiating their contracts for next year, which data set is most important? Data Set "B". Now as the teacher in the classroom, you always have to keep in mind that those won/lost records, those summative scores that appear in the paper from high stakes testing, in California we have something called the API, the Academic Performance Index, which is another piece of political

data. It's still important, it controls a lot of things. It controls what the public thinks about you because your public cannot get into this data set or would they necessarily understand it if they did. So it's something you've got to keep in mind. But it helps you keep it in perspective as a teacher, because I know myself as a teacher sometimes the political data or the data that judges concerns me when that's what's seen. This is a good metaphor though; this is a good analogy for explaining how that fits in the world. What we're going to try to do today is see how we might be able to take Data Set "A" and by slicing it and dicing it, turn it into educational data that you can use in the classroom. Okay, the second thing I want to talk to you about from an analogy standpoint is the difference between Criterion Referenced Tests and Norm-Referenced Tests. Ten years ago, most every state in the union, if they were using any type of statewide assessment, or the districts that were using district-wide assessments, were using Norm-Referenced Tests. Norm-Referenced Tests are built to accomplish a specific function. Now, let me give you an example, and what's happening now is that we're moving from those Norm-Referenced Tests to Criterion Referenced Tests, so I want to give you an example of a Norm-Referenced Test from a very hypothetical situation. Sandy, me, I'm going to race a hundred meter dash, and I'm running as fast as I can and there are a hundred people in that race, myself and 99 others. I take off, I run as hard as I possibly can, I cross the tape at the finish line and I come in last. Tell me, this is not rhetorical, please answer, tell me how I ran. Come on. Slowest, see now the trouble is you all are educators and you're thinking. If I do this at just a random group, I will get "don't quit your day job", "you're so slow, why are you even out there" things like that. Okay, the point is, as somebody pointed out back there, the only thing you can tell me is that I was last out of 100 runners. How well did I run? I was last out of 100 runners. That is a Norm-Referenced Test. When you take one of the large nationally Norm-Referenced Tests, you are comparing your kids to a norming group, that was a snapshot taken, usually across the country, with just the right amount of ethnicity, poverty group, English learner group, etc. across the country into an representative sample and then your kids raw scores are compared to the raw scores in that norming group to determine what percentile these kids fall in. It compares, but it compares based on that national norm sample. Okay, let me tell you about that same race, but let me give you some more information. Those were the 100 fastest runners in the United States. These were the Olympic trials for the finals. The top three runners from this race were going to go on to run in the 100 meter dash in the Olympics. Here's something that's most important though. My elapsed time, in running that race, was 10.20 seconds. Now tell me something about how fast I ran. Wow! Fast! But most importantly, from a teacher, from a coach standpoint, you can tell me that I ran 10.20 seconds. You have a precise criterion to evaluate me on. Right? Now, look at this, this is a Criterion Referenced Test. I was running with a lot of high paced people but it didn't matter. You're not comparing me to them now, you're only comparing me to the criterion of what time does it take me in seconds to go from the beginning to the end of that particular race. This particular type of comparison has no control on a bell shaped curve. It's not influenced by and it's only confluence about the criterion in you. Now, how well did I run? I ran 10.20 seconds, 10.20 seconds is world class time. It was world class time. But it is possible, in the United States right now to run 10.20 seconds and still come in last out of 100 runners because we have approximately 100 people in the United States who can turn that kind of time right now. Okay, think about this as a teacher. Which is more important for you in the classroom? Norm-Referenced Results or Criterion Referenced Results? And that's where No Child Left Behind, the requirement for content standards across the states across the nation, that's where we're being influenced to go. And those are the type of assessments that

come out. Now they're usually not as stable as Norm-Referenced Tests because with a Norm-Referenced Test, you're compared to the same group over and over again. If you're comparing state to state, those are very effective instruments and they're very effective for a lot of other reasons, but when you're talking about teacher level data, Criterion Referenced Tests are extremely powerful. They're extremely powerful tests but they don't come with the same lingo attached to Norm-Referenced Tests. You don't get percentiles because you are not comparing. This is important and the interesting part is it was the same race. It was just analyzed two different ways. In fact, the big testing companies who do Norm-Referenced Tests will also give you Criterion Referenced Data very often. Okay, just to put this in perspective. Jesse Owens, for those of you who were around in 1936, turned 10.20 time to set a world record at the Berlin Olympics, but right now Tim Montgomery holds that record and if I could turn that 10.20 seconds, which is about the time it takes me to get to that door in my present physical state, I would be one of the top runners in the United States. Okay, I want to introduce you to the vernacular that I will be using throughout the rest of this presentation. When I refer to curriculum, and this is what we came up with in our research in Riverside County. When I refer to curriculum, I am referring to the state content standards. In other words, the use of the word curriculum for this presentation means the what. What are we required to teach? Now this is the minimum core curriculum. This isn't all the other stuff you do, and all that stuff is important, but because of the time restriction and because of where this presentation is going, curriculum is synonymous with the content standards for the core subjects. Instruction is the how. Curriculum was the what. Instruction is the how. Instruction really is made up of three elements, the teacher-of course the most important element, the instructional program-the textbook or whatever system you use, and the process that you use to communicate it back and forth to the students. That's the how. Now, learning is the why. That's why we exist. The learning is the student assimilating the knowledge and values from the instruction side of the house, mainly the teacher. Now here's the question and here's the reason that I exist. See this red arrow? Now, answer this question please. Because you teach it, will they necessarily learn it? No. Do I have any yeses? Okay, because you can leave. Because you teach it, they don't necessarily learn. They usually do, but if we can have the best teacher, with the best instructional program and the best everything in the world, and the best student and there will be times when learning does not occur. There will be times when learning occurs one day and it's gone the next day and then it reappears later. There are lots of various situations. But this red arrow is critical because if you can measure the difference between what you think you are instructing and what the students are actually learning, you are miles ahead in the process of dealing with the data that falls out the other end. Now this sounds so simple, but remember we spent four years working with teachers all around our county in order to determine the best way to get them data. And this was a labor of love, believe it. Measuring is what myself and the people I work with do. Measuring is what the teacher does. Measuring is what we all do. If a carpenter is cutting a board to put into a wall of a house, the carpenter will normally measure the board, mark the board, cut the board, measure the board again to be sure, and then install it. If it doesn't fit, something has to be done. Okay, in this case if you can measure the difference between instruction and learning, and we call that assessment notice I don't call it testing, to us and to this project, test is a four-letter word. A test judges, measurement informs. Test is the judgment or political data. Measurement is the informational/educational data. If you can accurately measure that difference, then you can properly design an intervention to fix the instruction so that the students now learn what they did not learn before. One of the critical elements that came up in our research, in Riverside

County was is there any way we can determine why? As an assessment person, I taught, I administered at the site level and I administered at the district level. When I administered at the district level, my primary job was assessment. I was very proud of myself when I could bring to a teacher a pile of reports, usually delivered with a forklift, that would tell you in every possible format available, what you're kids don't know or what your kids didn't know. And you would look at me and place it over there with some of those papers that hadn't been graded in 6 months. I know that doesn't happen to you. However, the point is that what you really need to know is not only which kids don't what, but more importantly, why they don't know it. Because if you know why they don't know it, hopefully, you will be in a better position to plan the intervention. Okay, think of these three terms throughout anything I'm saying, Measure, Analyze, Intervene. You want to measure the difference between instruction and learning. You want to analyze the results to try to figure out what the kids don't know and why they don't know it. Then you want to intervene in order to correct the problem. Testing and Measurement, one more time, because this is near and dear to us and something that we found out was really bothering teachers as we were doing our research, testing judges, measurement informs. If I were back in the classroom, my first year teaching in the public school system was in fourth grade. If I were back there again to that Friday spelling test, with twenty questions on it, I would approach it a completely different way than I did twenty years ago. Now I would talk to my students instead of saying "well, we're going to take that spelling test tomorrow and if you get 18 of them, you'll get an A and if you get 16 you'll get a B and if you get whatever else you get a C or a D or then I'm going to call your parents" and stuff like that. The way I would approach it now is "Tomorrow we're going to take a spelling test. This week we've been working with 20 words, our vocabulary words, we've used them in sentences, we've talked about the etymology, we've moved them around, we've put different prefixes and suffixes on them, we've written with them. Tomorrow we're going to determine how well you know how to spell them. And the reason we're doing that is because I'm trying to find out how well I've taught this. I'm trying to find out the difference between what I've taught and what ya'll learned. Ya'll help me and do the best you can on this and then we'll analyze these results come next Monday and we'll look at what we can do to shore up any problems or celebrate the results, depending." During the course of the four years of this research we've been doing, we've had actual teachers try this particular technique in their class and the results have been astounding in the fact that the students have become much, much more desirous of being assessed or measured than they were at being tested. Not a technical definition, but a very practical one. Four essential questions, you've probably heard these and have heard them in several different presentations today. What are students supposed to learn? That is defined at least at the minimum core level by the content standard, both for our state accountability system and for NCLB it's probably the same way with you. How do we tell whether the students have learned what they're supposed to learn? Well, that's done with measurement and the rest of this presentation is about that. Here's the third question that wasn't there before. How do we tell "why" the students are not learning? That's this process I'm going to talk to you about that we developed with the teachers in the county. And then, what do we do to reverse a non-learning situation? And we call that an intervention or remediation. I'm going to ask you three very, very simple questions and the purpose behind these three simple questions is to show you how simple the actual answer to our problem was. Now the answer was simple. The implementation was not so simple. The other thing these will help you do is to deal with your students with regard to multiple choice type questions. Okay, first question: We're preparing to run a marathon in twelve months, what are we going to do to help us prepare to run?

First, shooting free throw shots, second kicking field goals, third practicing penalty shots, fourth running 100 meter dashes, or fifth long distance running. Now, I'm going to count to three and you give me the letter. One, two, three... "E". Anybody pick anything else? Now I was told, that's the right answer, I was told by some teachers that were sitting here in a previous presentation that the best answer is really "D" and "E" both of those. People who know something about physical fitness and that is true, but if this were the question on your state's high stakes testing, you better have the students read all of the answers and choose the best answer. Okay, this is relatively simple right? The next one is a little bit harder. This is something I actually tried once. Now, I live in southern California and the objective here is to sail a 32 foot sloop around the world with two people aboard. I got from Newport Beach to Catalina. For those of you familiar with California geography, that's not a long way. But what would you do to best prepare you for an endeavor such as this? Crewing on a 12-meter yacht, attending open water survival training, participating in catamaran races, entering the local speedboat competition or gaining proficiency in sailing a 32-foot sloop. Now, One, Two, Three... "E". Imagine that, you're much better than the last group. Okay, probably "E" and "B" if you are able to choose two. Now, let's bring this home to where it really matters. You are teaching your students in accordance with the state content standards. We all do that now. Now if you want to measure their progress along the way, what's the best way to do it? Assessments from the text, Assessments from supplemental books, the ones you go and buy at the local teacher store, Standards-based assessments that are closely aligned to the content standards, or measurement items you make up the night before. One, Two, Three... "C". I had a couple of others here but I think the majority of you said "C." "C" is probably the correct answer, but I will tell you, as a teacher who taught adults and kids and teachers that I have done all four of these and I bet you have too.

(MUSIC)

Sanford - Segment 2

(MUSIC)

What is the only problem with using assessments from the text? I'm in California, that's where my schools are and that's where the people are that I work with every day. California has more students in it than any other state in the union and we can't get a textbook publisher to write a textbook just for our content standards. They get close sometimes. So, what chance does any other state have? The point being, you cannot make the assumption that the textbook is nailing your particular state's content standards. You can't make that assumption. Used to be, you could teach from the text and let it go at that. You cannot do that anymore. Your state content standards are the authoritative source. The text is really a resource that supports that and you have to keep in touch with those standards. And in fact, if you're really a superb fourth grade teacher, you're not only in touch with the state content standards for fourth grade, but you are for third and fifth grade as well, in order to form that vertical articulation that is essential if you're really going to take off into good, great. Okay, assessments from supplemental books, I will tell you the reason I used to do this, if I had kids way below grade level, I would go out and get something to assess them with just so they would show some positive interaction with the actual assessment. That's fine and dandy, but don't forget when the high stakes testing comes at the end of the year, they're going to be held responsible for the standards that are there. And then measurement items that you make up the night before, we all do this, but when it comes to

exactly aligning with the content standards, it is tough because of this. There are three required criteria. You've got to align it with regard to content, but you also have to align the item with regard to skill level. In California, in the second grade, there is a content standard for synonyms and antonyms. There is also one at twelfth grade and every grade in between, but the skill level required at those various grade levels is considerably different. At the 2nd grade it's essentially identification of synonyms and antonyms. At that 12th grade it's metaphorically using synonyms and antonyms in creative writing. So you've got to be really careful that as you're graduating up between 2nd and 12th grade in English language arts that you nail that exact skill level that's supposed to be there so that you're evaluating what's going on with the kid. That is not easy. Isolation is not easy either. That means not assessing any standard that absolutely isn't essential for the standard that you're trying to assess. It's not easy. Now the one additional one we came up with after working with the teachers is designing the assessment. Since most of these assessments that we take for high stakes testing are multiple response items – multiple choice items in common vernacular. What the teachers needed – what the teachers wanted – was to make the wrong answers as powerful as the right answers. Give us wrong answers that point us towards the kids are probably thinking, instead of just random choices that are incorrect – and a great deal more about that in a few minutes. Now for the first activity – and these activities are not meant to take a long time – here's what's going to happen. In order to show you the evolution of benchmark assessments and assessments in the classroom that we've encountered in Riverside County – and we're pretty much sure that the whole country has been in some way, shape or form in about the same shape. I'm going to show you a copy of a report that 5 years ago – it's names have been changed and the report has been simplified to show on the slide – but this particular report was used to give information back to teachers once a benchmark assessment had occurred. And what I want you to do at your three tables is to discuss two things. First I want you to discuss what aspects of this report are not useful. And when we finish the activity and I ask you I want you to give me one or two. And then I want you to tell me which aspects are useful. Now let me give you an example for this first go-around. This is Mr. Smith. This is Mr. Smith's room. Mr. Smith got this report back after a benchmark assessment. The top portion up here lists all the students in his room. He has more than 4 – don't be jealous – I just left those other ones out. Let's say this is a 4th grade math assessment. It shows his class average. Those figures show the percentage correct on the assessment. At the bottom it shows the results for the grade level as a whole. Say there are four or five 4th grade teachers at this particular elementary school. Now what is supposed to happen is the 4th grade teachers are supposed to meet. They each have the results and they have the results of the grade level and they're supposed to go over those and determine how they can improve mathematics instruction based on this report. Is everybody with me? What I want you to do is take 2 or 3 minutes – don't be exhaustive – but think about what about this is useful for your mission of trying to improve mathematics instruction, and what about it isn't useful to you. Go ahead. Overall we have a pretty good idea that they're meeting standards overall. The problem that we get into is that we don't know which part of the standards in math – which ones they're getting wrong – which ones they need improvement in. So we need to have an assessment that breaks it down. We can't change anything with the average score because we don't know what specific areas of difficulty. Exactly, it doesn't reference any content, specific information that you can use to figure it out. The only thing we've gained then is we know that Johnny needs lots of help. We also don't know if the students are making mistakes on the same items or skill or a variety. Table 1 – tell me something about this report and the data it contains that is valuable to you in

your mission to try to improve 4th grade mathematics instruction and student achievement. They're doing okay in math. I mean it tells us that they're not way far away from what we're trying to do. The problem that we get into is that we can't use this to drive our instruction. It's not telling us where we need to improve, what standards that we need to match, and it's not telling what areas we've done well in. And that kind of summarizes what the rest of you said. Let me show you – we actually took these reports and went through and surveyed hundreds, thousands of teachers to talk about these exact issues. And the teachers said – and this was after they were in a standards-based environment – they're in a standards-based environment, they're meeting as a grade level team to improve instruction, but really to improve learning – in other words, to improve the instruction for the purpose of improving the learning. And they're in a standards-based environment, so they need data at the standards level. They need to know which standards are getting. And there's nothing in that report that gives it to them. But there are some things. They were getting something. Many of the teachers we talked to complained insistently about the fact they would benchmark tests sometimes and not get the results back for a month or two or something like that. Okay, there was some quantitative data there, and as was pointed out by several tables there were ways that you could look at that quantitative data. There was student-level data. In other words it wasn't just for the classroom as a whole. But most of the results that came there were summative. It was giving you a percentage of items correct for math – and math is pretty broad. And in 4th grade in most states you're going to have anywhere between 20 and 40 content standards for that grade level for math – so it doesn't point you. There's no indication as to what the students don't know – only indication as to where they're scoring on a scale. And no reference to strands or contents standards, and I'll talk more about those in a second. Another thing is we couldn't tell how long it took to get the data back because the #1 complaint from teachers in our first wave through this program – and this was 4 years ago – was get me the stuff back sooner. And no hint as to why the students didn't miss it and you've pretty much covered all these in what you were saying. From the first round of surveys though is the teachers need to get the data back quickly. And they not only need to get the data back, but they've got to have a copy of the assessment. Secret assessments don't work for this purpose. In other words, if the district or the principal or somebody comes down and administers an assessment – if you're going to use that later for analytic purposes, you've got to have a copy of the assessment. Ideally, or optimistically, the teachers wanted it back within 1 to 3 days. They figured that if they got the data back within 1 to 3 days they could deal with it. Much longer than that, the instruction is starting to pass where they were with the data. And ideally they wanted it instantaneously. So we worked off that data and we went back and we started looking at something a little bit different. Now this statement by Jim Popham, who is a nationally known program evaluator and testing person, says that in a standards-based environment, if you do not have data at the standards level you are lost. If I am holding you responsible to teach certain contents standards, unless I can give you a mechanism for measuring the student achievement in those contents standards all along the year – whether it's with formal evaluations that the district administers, or informal each day as you go – then you cannot properly do your job. Okay, the vernacular that I'm using here is a content area, like English language arts math – a domain we use primarily in English language arts, there's a reading and a writing domain. And a strand is a set of standards that have a common characteristic – for instance there might be 6 or 7 standards that make up measurement in geometry, or reading comprehension, or something like that. And the most granular level of measurement being the actual standard itself. Okay, we're going to do the same thing now with the next level of report. Now this next report was one that in some way,

shape, or form, we were commonly administering 3 years ago. We were commonly using this report to give information on data that we had been measuring about 3 years ago. And what I want you to do with this is the same thing you did last time. What aspects are useful and what aspects aren't useful. Now look at this particular report – and to avoid confusion, because it is confusing – but I'm giving you exactly what we were giving the teachers at the time. And it's abbreviated. This is an English language arts assessment report. There were 30 questions on this particular assessment. They were in 2 strands – written and oral language conventions, and writing strategies – so there were a number of standards in each one of those strands. And then for each teacher, they get the names of all their kids in their class. They get a raw score that says how many questions they got correct, and then a percentage – not a percentile, but a percentage. 20 equated to 67% overall. And then it also breaks it down for Johnny, Mary and Ted and the rest of the students by separate strand, so they do have more granular information. And then it also gives the class average. And you can get this same information back for grade level, for class, or for individual. So it gives it all at those different levels. I think basically you have more numbers to deal with – in other words you have more data so you can make a better decision. You know how many questions were asked, and I think that's important, because otherwise if it's 3 questions what does that give you to choose from in terms of data. And then percentage correct is critical, because you want to know where the gaps are for your class and individually for the students. Okay, what about this is useful and an improvement over the previous report. It's better because it's starting to actually break down the different types of strands – he's feeding me words with my limited vocabulary. But it's still not taking it quite far enough, because you don't know exactly what each kid has answered on each question. So you want – in the measurement game – what was call more granular information. Now notice – I'd like to point out – in 1999 to 2000 when we were basically using the first kind of report, the teachers said we want something that's more granular and we gave them this. And then what's the next thing that came out of the teacher's mouth? We want something even further, and this continues to be the case. And this is a work in progress. We've figured out – we're never going to get to the end. Hopefully it will always get better, and better, and better. But some of the school districts we work with will hand them one thing and they will say this is great, but 2 weeks later will say – but, but. If you could just put this button in there. So what you're talking about is what we found out was the natural evolution of the desire for more data. But notice the difference in not using the data and desiring more and better data. This is where we begin to think – this was the stage we were at about 2 to 3 years ago – maybe we're making progress. Okay, let's go on and go to the next phase here. Before we start – and let me go through these – a word about the level of granularity. You name is? Dara. Can you tell the difference between Dara and myself? Okay, if you were just looking at our heads could you tell the difference? If you were just looking at our hair could you tell the difference? However, if I was to remove a hydrogen atom – and there are many in Dara's body or mine, which there are many – and I was to place each of the hydrogen atoms side by side on this table and you could see a hydrogen atom, could you tell the difference between us? No. The point being is there is a point at which you can take the data apart too far, where it becomes meaningless. And it's almost hypnotic, once you start taking it apart, to want to take it apart further, and further, and further, and further. And it is an art to decide where you've gotten to the right point. So it's just a consideration to keep in mind. When we took that previous report – and that previous report is pretty much what, up until about 2 years ago our high stakes testing report looked like in California. That was the way the data came out of it. Much more, but basically in that type of format. Results were

forthcoming. These were the positive things that teachers said about it. A grade level summary provided, a disaggregation by strand – and this was the important thing. We had moved from just a summative evaluation of math or English language arts, and began to take it apart by pieces so the teachers could say where do I have to put my effort. Remember, what is your most valuable resource? Your children – the ones you teach. But what is your 2nd most valuable resource? Time – there is never enough, nor will there ever be enough time to do everything you need to do. So anything you can do or analyze data-wise to give you an advantage over the element of time is a great advantage to you as a teacher, and more importantly to the student as a learner. Classroom level data for teacher by strand, student-level data – you can see we got almost twice as much as we had in the last report, but they were still getting them too slow. In high stakes testing in California, tests are normally given at the end of April, first of May – around that area. And they get the results back usually in August or September. By that time you've got the next round of students, etc., etc. Okay, no indication as to what the students did not know. No indication as to why. So what were we able to conclude from that? Because that last report format was basically a high stakes test type situation, you cannot rely on high stakes testing data. Those tests that you take at the end of the year. You cannot rely on them for your exclusive diagnostic data for the next year. For one thing it's different kids. It can still give you something about your program – where it's high and where it's low in a general sense. But it does not give you data usually in most states at the standards level. It gives you data in a broader format, and you need to tunnel down – to drill down – to that standards level. Okay the next one – slightly different request here. This is the report that we were using as a report format that we were using as of last year for our benchmark assessments in most districts in the 23 districts that make up Riverside County. Riverside County has about 330,000 students – 23 districts. What aspects of this report are not useful? And then, what would you want if you could have anything you wanted – anything you wanted pertaining to this? Don't ask for a Lexus? Okay, now, this is abbreviated, so fill in the blanks. In this particular report, notice that at the top we have the strand data, much as we had before. And there might be many more strands than just these strands – 30 questions on this particular assessment – 30 items. Over here, this is a list of all the items – every one of the items – from 1 to 30. I've just put 3 of them up here. So the 1 and the 7 and the 13 were selected just random. And then there's the percentage correct. Now remember, this report – you always want to start analyzing your data at the grade level, or the content area level – at the largest level possible, with the largest number of students in the sample. For a statistician – for a measurement person – there is power in numbers. If you looked at this – you were, let's say in the 4th grade, and you had one classroom and you looked at your data there first, you'd be looking at approximately 30 give or take students. If you're looking across the grade level, you're usually at somewhere between 80 and 120 or 140 students. If you analyze your data first, and not even look at your classroom data, you will pick up trends in the instructional program, the way the instructional program is administered and how it affects student learning. And then when you go back, after you've looked at that, after you've thought of interventions, and you go back to your individual classroom, what you'll find out is that the number of students you have to deal with still is much less than it would be before, and then you get down to the actual student level if need be. But for each one of these we have percentage correct, percentage incorrect, percentage that didn't attempt it – that's a telling statistic sometimes – we have the standard number, but then we have the noun description of the standards spelled out. When we first tried this form of testing we just put the standards number there, and in our state then a teacher had to pull open a book and look up the standard. Well, our

standard numbers aren't even unique per grade level, so that caused problems right there. When the teachers talked to us, they said Sandy give us something that has all the logistic work done for us so we can spend out time figuring out what's wrong and what we have to do. So, every time that we show a standard now, we spell it out. It doesn't have all the words, but it has enough words to where the teachers know what it means. Now in this particular report you can pull this report up by classroom, by student, by grade level, by content area – secondary for algebra I or chemistry or whatever – or across the district at any one of those particular levels. So right now perform the activity again and come up with what's still not there? And what would you like if you could get anything you wanted. If you can get it by student and classroom that's helpful. As a classroom teacher you'd want to know which things you're doing well and what the kids are picking up and what they're not within your own classroom. What I don't know is what is that standard, is that like in Wisconsin when we looked at our WKCE results, it's 4th grade standards. But then in 3rd, 2nd, and 1st, we use that standard to guide our teaching all along the way up. But I don't have a test in 1st grade – I don't have a benchmark test that says by the end of 1st grade they should be able to do this, because in 4th grade it needs to be mastered. So I don't know what that standard number is. Okay, let's wrap it up. Table #2, would you like to report? We said the one thing obviously that we want to know next, is well, okay so I have only 23% of my class getting grammar correct. Grammar – it doesn't show the rest of it, but it would. So then now I need to know what do I do about it. If this is 4th grade, what do I have the 3rd grade teachers do to prepare them and what do I have the 2nd grade teachers do, but that would be part of that. Well, think out of the box. That's what we're looking for here, and that's exactly what our project did with the teachers in our county. So now what I would like as a benchmark that 1st grade should have so that they can reach the proper benchmark in 2nd grade so that by the time they take the 4th grade test they're able to reach the benchmark. And in the vernacular we use that would be full visibility. You'd like full visibility over everything that's going – there'd be no secrets. Nobody would be trying to guard their stuff. Everybody would be sharing their data, because the common goal is the increased achievement of everybody. Remember back to our report two reports ago, Mr. Smith – he was scoring about 3 1/2 percentage points higher than the grade level as a whole. Does that really mean that he's a better teacher? Not necessarily. He could have all the high kids – there's no way to tell from the data that was on that report. We also talked about the power of this when you're meeting collaboratively with an entire grade level team to really have a conversation about gee, on question number one, I only had 23% of my kids get that, but then had 70% of his kids. Let's look at how that was asked and Vince how did you teach it. I used this lesson. And then he might say well wow, based on the pre-assessment data I had I really tweaked that a little bit and this was effective, and then you've got the teachers really collaborating together. Visit – maybe going to observe each other teach certain components and then they're improving their instruction. In order to do that – and this brings is kind of like a cause and effect working both ways. If you get that type of collaboration, all of a sudden ego goes out of it. It's not who is the better teacher, because it's about achievement now. It's about student achievement, not about teaching. It is about teaching, but about teaching that raises student achievement.

Music

Third segment

Music

Let me move right along here. I don't want to get too far behind. These were the kind of things that the teachers we talked to said "This is really good. The phenomena occurred. We really like this. Now we've got data by standards. We can really use this. But there's a list of 18 other things we'd like for you to put on this." We could not develop – and we're still working on this – we can't develop stuff as fast as people want it. And that's the right position to be in. But they want to know why and that's what I'm going to get to now. And then what do you do to fix the problem, and how thoroughly the standard was taught, and how long has that one been out there? Because if I'm going to assess how well the students did on it, the big difference is if we've been working on it 3 months or 3 days. Save teacher time – this is where it came up – hey all this stuff you're telling me is nice, but don't give me something else to do. In fact, take something away from me so I can really spend some time playing with it. And tell me where the standard is taught in the instructional program. Here's the summary of what we found from 4 years of working with this – timely summative reporting – all these, but these are the ones that kind of jumped out and surprised us, because they weren't there at first. You know, have full sorting accountability, pacing stuff, all that kind of jazz. Okay, here was our solution. We had to have a bank of items, and it had to be big, with items that could be used, multiple response items, open response items, all kinds of items that teachers in the classrooms could use to assess student's standards – standards for their grade level and for other grade levels, and for all subjects. And, the multiple choice ones were our first target, because they're the ones that can be analyzed quickest. Then there had to be a process to do all this stuff we've been talking about. And then we had to have all kinds of software support because you can't do this manually. You can't sit down and score a test standard by standard for every kid in a grade level without it taking you an extraordinary amount of time by hand. Okay, so together we invented a process we call WHAT – what and how analysis for teaching. And when we're giving these presentations inside Riverside County we call it what's on first. And it's based on these 5 things; Teacher-friendly data, top down processing. We talked about teacher-friendly data, top down processing. It's basically analyzing at the highest level first and then working your way down. Wrong response analysis, which I'm going to talk to you about in a second. Pacing, which I'm going to talk to you about and the supporting software tools. Okay, top down processing simply means this. It's best – don't even hand the teacher reports out until you talk about the grade level as a whole, or the subject area. And then while you're at the same meeting, then hand out the teacher reports and start talking about these things table #3 was talking about with regards to what did you do to get that? What did I do to get this? And of course the next incantation of this is when you're doing it back and forth between grade levels. That vertical articulation is extremely powerful and probably the highest level we can aspire to in our time of really getting that type of cooperation. Now, look at this item. Look closely at this. I'm going to read it in case you can't see it. Last year the cost of tuition for one semester at a certain college was \$1,200. If the tuition was increased by 20% what would the cost be? This is a 6th grade standard in California. I think it's a 5th grade standard too – somewhere around that. Okay here are the 4 possible answers. Now C is correct, I'll tell you right off. I'm not asking you to work the problem. I wrote this item. You'll notice it says \$1,200. I wrote it in 1919 – that's when you could get a college education for \$1,200 a year – I'm older than I look. The first important part here is that each one

of these distracters was not chosen by chance. And in fact it can be done in different ways. In this particular case, this particular item was administered as an open-ended item to 400 different 6th graders, and then the items that came in from the open-ended part were analyzed. Now you see in this way we are helping the teacher, because we're giving you the convenience of multiple-choice question with the power of an open-ended question, which has always been one of the problems in assessment. And ideally – and we sift these things through several cognitive psychologists we've got and through teacher teams to say what is the most likely cognitive disconnect for a kid who is in the 6th grade who has been exposed to the standard, but for some reason can't nail this question? Now, this is normally the way you would get the results. This is the traditional way. Okay, item #5 was on computation of percentages, etc. and 63% of the students got it right. What good does that do you as a teacher? What good does it do you? Answer me, please. What good does it do you? Let's get angry! It does not do you any good. Because 63% got it right – that's some information, and it's better than you used to get, but that's really not what you need with what we're dealing with now. How about this? First, you have the item. And then you have the frequency distribution. Remember we're looking across the grade level, so this is a 6th grade item, if it was a middle school it could be 500 or 1,000 kids. If were an elementary school, which some 6th grades are, then it could be maybe 140 or so kids. But what can you tell immediately about that now that you couldn't tell before? Where are the majority of the kids missing it? And why? What does "A" tell you? Inaudible. Yes, what the kid did is, and here's what we went one step further, because we get this result. This is the answer you get back. They computed the 20% but did not add its original cost, or – this is abbreviated – or they didn't finish reading the question. Now with this information – and this is actual data – this is actual data from a test school – how long do you think those teachers said it would take to invent an intervention to fix that problem? But if we look back – if all you got was that, how long would it take you? I don't know if you could. You don't have the information. So this was the big step. This step was crossed about one year ago and we've been working like heck ever since. We could not buy an off-the-shelf bank of items. So we've been writing them ourselves. And that is not easy – that is not easy. We have professional item writers, we have teachers working with professional item writers, we have cognitive psychologists working with them, and it's a slow arduous process. Nowhere quick as we wanted to, but they love this. They love this. Now let's carry it a step further. Okay, that is what I call a wrong response analysis report that you just saw, because it concentrates on the wrong answers, provides rationale for incorrect responses – that gives you a leg up. Provides hints to the cognitive disconnect because it provides that. And here's something that is near and dear to us who really know – and that's all of us – who really know what it's going to take to get the students to the highest possible level of achievement. They have to be aware of the metacognitive process. They've got to think about thinking, because what is most important – teaching a kid what to think or teaching a kid how to think? If we want good citizens, if we want to make progress, if we want to be competitive in the world, and at some point the universe, it's how to think that's important. And when those kids are analyzing the results of those tests themselves, which some teachers are doing, they're going through that process. Why would little Johnny choose? And it's not Little Johnny, but why would any student choose this? And they'll work it out and they'll figure it out and they'll be taught how to think. Here are some other ones. Here's a 1st grade item that we field-tested. What can you tell right away about this? Now, in California the kid is not expected to read the word contraction. We're not that far advanced. However, this is an item that is read to the kid. The kid is expected to understand what contraction means. Well, when the teachers

in a certain school gave this, they knew right away that 42% of the kids did not have the most basic understanding of what a contraction is, which is the presence of an apostrophe, so that helped them. Everything is harder in high school. Look at this. And how many of you would bet the car in the parking lot that you can nail this one? Okay, now the advantage of this process is that as the teachers are going through it they're analyzing the results of this. They get a frequency distribution and all of that, but they're able to get the rationale for each one of the distracters. So the point being for this particular stage that there is staff development taking place with the teachers as they're going over the analysis, as they're going over each answer, because they're learning what each of those other words besides consonants, which was the correct answer, means. Impressive? Yes, I think so. Simple? Relatively. Has it been done? No. Okay – pacing. This was a big deal with the teachers after we got them to a certain level. If we're going to assess each standard, some standards are taught at the beginning of the year. Some standards are taught at the middle of the year. Some standards are taught one month before the assessment. Some standards are taught 5 months before the assessment. How do we tell the difference? Well, we pace them. In other words, there's a color code assigned that when you see the report gives an indication of how long that particular standard has been out there and active. Okay, J.D., would you do me a favor please? The last part of this presentation is going to be a demonstration of the software that we are now beginning to use in Riverside County to accomplish these goals. Is it perfect? No. Do we still have bugs? Yes. Will we continue to have bugs? Yes. Will we fix them? Most of the time. Will we run into roadblocks? Of course, but we will continue to listen to the teachers, because the next time we have them back they'll want 10 times more than they have now, and that is good. Okay, this would be equivalent to the computer screen if you'd just done that 6th grade English language arts assessment and the 6th grade teachers were around the computer screen or whatever medium you wanted to use. And the first you'd normally do is click on that top one and you'd get a report. You've got a copy of it I've given you in color that looks like this. Now what this does is list every item on the assessment in order. It lists it in order from top to bottom. Now the reason you want to do this – it hasn't aggregated them by standards, because remember there are multiple items in each standard. But you're looking at it like this because the first thing the team wants to do is get familiar with each item – understand it – see if there are items in there that are incorrect, or maybe bad items – that happens, because you can touch a button in the software and take those out of the system and it automatically reports back to us and tells us that there is a problem with an item if that occurs. So it normally takes you just a while to go through these and you'd have a copy of the assessment itself and you'd be looking at each one of these and seeing how they check out. Now on the software you can click on the top of each one of these columns and it will sort by that column. You can click on certain places, like right over here and it will give you the list of all the kids and how they scored for that particular situation. Now click back to list over here and so after you've spent a little bit of time on that you'd click on the one that says “by standard”. Okay, now this is the seminal report, because what this has done is it has aggregated the individual items by standard. So, look at the top one up here. This particular standard has two items that make it up. This is a very abbreviated one it could have many more. Now what catches your attention as a grade level team, because this particular report is automatically sorted from most missed to least missed standard. You can re-sort it any way you want to, but it's automatically sorted most missed to least missed – that's the default. And you'll notice this top one was aggregation of these two items for that particular standard 92.31% of the possible responses were missed. What attracts the team about this? Yes. This is a red one. Red ones

should have been mastered. Ideally the top one, when you sort the report like this, would all be the standards you haven't covered yet if you decide to put any of those on there. And then the ones that you just introduced, then the one that you practiced, and at the very bottom would be the ones that should have been mastered. But does this kind of thing happen? You betcha. And we've found it does. And the teachers, instead of being angry about it, were extremely happy – you know something's wrong. We're teaching it, but they aren't getting it, so let's figure out why. So the next thing we'd do J.D. is click right on there. It opens up on the computer screen and it breaks it down by item. So the items over here come out – well this one item incorrect 96.15 and this one 88.46 – okay therefore it's not one of the two items, it's both of them. Okay, the next step is – let's get serious – click on results for question #6. This comes up as a detail screen. The teachers are looking at it and the first time they saw it they had their mouth open, just like you did. Wow! Okay, there's the question at the top. There's the item – what we call the item stem. And then here is the frequency distribution, the answer, it shows you which one is the correct answer, and the rationale that would be used for selection of that answer in place of what should have been the correct answer. Okay, now you can see here in the 2nd item – I'm not going to take the time to show you the other item, but it comes out about the same way – what's happening here? The correct answer is simile, but these kids were jumping on hyperbole. So they're getting the words simile and hyperbole mixed up. How long is it going to take you to remediate this? Not as long as if you didn't have this data. Now here's the really cool part. If you can click over it will take you to kind of a look-alike of the scope and sequence for your particular instructional system with a little arrow showing you the best place to go to get materials for intervention. Now, is this an improvement over what is being used in your schools now? Is that universally yes? Then the applicability of what I'm talking about here for you is exactly that. It's a different way of looking at data. And the mechanics of all this are nowhere near as important as the concepts. It's a different way of getting data to the teacher level so the teachers can make adjustments in the instruction to improve student achievement, which is what it's all about. High stakes testing – that's a by-product – that will take care of itself. You don't have to teach to the test – you teach to the standards – and if you do that the testing will take care of itself. If you'll hit this one over here please and then just click down here somewhere. If ya'll want to know anymore about this, these are the people I work with in Riverside County. You can e-mail any of us anytime you want to and we can talk to you more about this. Thank you very much for your time.

(MUSIC)

For more information or a free online follow-up to this program, log on to [www.ed.gov/teacher initiative](http://www.ed.gov/teacher_initiative). This broadcast and the follow-up are brought to you through a partnership of the U.S. Department of Education and the Panhandle Area Educational Consortium.