

# **National Association of Test Directors 1998 Symposia**

Organized by:

**Maryellen Donahue**  
**Boston Public Schools**

Edited by:

**Joseph O'Reilly**  
**Mesa (AZ) Schools**

This is the fourteenth volume of the published symposia, papers and surveys of the National Association of Test Directors (NATD). This publication serves an essential mission of NATD - to promote discussion and debate on testing matters from both a theoretical and practical perspective. In the spirit of that mission, the views expressed in this volume are those of the authors and not NATD. The paper and discussant comments presented in this volume were presented at the April, 1998 meeting of the National Council on Measurement in Education (NCME) in San Diego.

The authors, organizer and editor of this volume are:

Sally J. Bennett  
Planning, Assessment, Accountability & Development  
San Diego City schools  
4100 Normal St.  
San Diego, CA 92103

Mitchell Chester  
Accountability and Assessment  
School District of Philadelphia  
Room 406  
21<sup>st</sup> Street S. of the Parkway  
Philadelphia, PA 19103-1099

Mary Lamping [Presented but no paper provided]  
Department of research and Assessment  
Milwaukee Public schools  
5225 West Vliet Street  
Milwaukee, WI 53201-2181

Ed Reidy [Presented but no paper provided]  
Pew Charitable trust  
1 Commerce Square  
2005 Market Street Ste 1700  
Philadelphia, PA 19103

H.D. Hoover, University of Iowa  
Lindquist Hall  
Iowa City, IA 52242

Maryellen Donahue  
Boston Public Schools  
26 Court Street  
Boston, MA 02108

Joseph O'Reilly, Mesa Public Schools  
549 North Stapley Drive  
Mesa AZ 85203

*Table of Contents*

**Accountability & Assessment: An Overview of Philadelphia's  
Accountability System**

Mitchell Chester.....1

**Revisiting The Issue of measuring & reporting Student Growth In  
An Era Of Standards-Based Reform**

Sally Bennett.....10

**Discussant Comments**

H.D. Hoover .....56

**Reply**

Mitchell Chester .....25

# Accountability and Assessment: An Overview of Philadelphia's Accountability System

**Mitchell D. Chester**

Philadelphia is committed to strong academic achievement for all of our students, and has adopted a comprehensive package of reform initiatives to reach this goal. While the focus of today's presentation is accountability and assessment, it is important to understand the depth and breadth of our efforts to raise expectations for what students and teachers can accomplish; help teachers deliver more effective instruction; support the non-instructional needs of children so they can concentrate on learning; focus resources, expertise, and decision-making authority at the school and small learning community level; and engage the public in shaping, understanding, supporting, and participating in school reform.

These capacity-building initiatives form the larger context within which accountability for academic progress is embedded. In fact, we waited to initiate the Professional Responsibility System until implementation of instructional supports (for example, the creation of the Family Resource Network, establishment of the Teaching and Learning Network, and implementation of full-day kindergarten) was under way.

### ***Philadelphia's Professional Responsibility System***

The goal of Philadelphia's Professional Responsibility System is to improve student reading, mathematics, and science achievement; promotion rates; and graduation rates. Our long-term performance target is that within one student generation -- 12 years -- 95 percent of our students will be proficient on system-wide measures of our standards, and will graduate from high school on time.

The Performance Index allows us to summarize a school's scores in reading, math, and science; its promotion or graduation rate; and the attendance of students and staff, into a single number. This single number does not provide details about the patterns of a school's performance or instructional program, any more than the single number of body temperature or weight gives us details about a person's health, nutrition, or exercise. However, like a temperature reading, the Performance Index is an indicator of whether a problem exists and whether things are improving, steady, or getting worse -- and, like weighing in, it is helpful in defining a baseline and setting targets.

### ***The Performance Index***

It is important to understand two things about the baseline (school-year 1995-96) Performance Index score and the performance target. First, we are aiming for the same 12-year target for every school in the district. It is not acceptable in the long run that children achieve at different levels depending on where they live, or their parent's income, or their race or background, or what language is spoken in their home. Second, while the 12-year target is a single standard, we know that we have to take into account where every school is starting. Therefore, the performance targets, which will be set for every two years, will vary for each school.

The Performance Index includes two domains. One domain comprises academic indicators, including reading, mathematics, and science, and promotion in grades

K-8 or, for high schools, persistence. The second domain addresses factors that help enable schools to improve achievement -- student and staff attendance.

***Performance Levels:*** The District has set a scale for each indicator (*advanced, proficient, basic, below basic*), similar to the scale used on the NAEP. However, to make the scale more sensitive, and to be sure we can show progress at the lower end, we have broken the *below basic* category into three levels (*below basic ///, below basic //, and below basic I*) We have also added a final category. In the case of test scores, this is for those students who do not take the test.

The *untested* category is included for two reasons. First, we believe that, as we seek to have 95% of our students achieve high standards, virtually all of our students should be taking the citywide tests. Some children -- for example, those who are mildly disabled, and those whose first language is not English -- will need testing accommodations in order to obtain a valid score. These students may need extra time, or may need to have parts of the test translated. Having the *untested* category gives a school credit for reaching out to test more students, because the Performance Index provides more credit for tested students than for those who do not take the test.

A second reason for including the *untested* category is that we know that the percentage of students who actually take the tests varies considerably from school to school, and we do not want schools that are reaching out to test all of their students to appear to have lower scores than schools where only the more successful students are being tested.

***Calculating the Performance Index:*** Each performance level is weighted when calculating the Performance Index. *Advanced* has a value of 1.2, *proficient* has a value of 1.0, *basic* has a value of 0.8, *below basic N* has a value of 0.6, *below basic JI* has a value of 0.4, *below basic I* has a value of 0.2, and students who are *untested* have a value of zero.

The index scores are calculated the same way for each subject area. Currently we are assessing reading, mathematics, and science achievement using the Stanford Achievement Test, Ninth Edition performance level scores of students in grades 4, 8 and 11. For any school and subject, the percentage of students scoring at each performance level is multiplied by the respective weight. By summing the scores for each performance level, we arrive at the Performance Index score for each subject.

Performance Index scores are calculated in the same way for student and staff attendance. To establish performance levels for attendance, we examined historical data for the district and considered what those figures would be a school where we would be happy to send our own child. For student attendance,

we set *proficient* attendance at 95 percent, which means that a student who misses nine days in a year is considered *proficient*, and one who misses less is *advanced*. Infrequent attenders are classified either as *below basic* /, if they attend between 10 and 74 percent of the days, or non-attenders, if they are on roll but attending less than 10 percent of the days. In measuring attendance, we are looking at all of the students in the school -not just grades 4, 8 and 11.

For staff attendance, we have also set 95 percent as *proficient*. We include in this measure not only the teachers and the principal, but all of the professionals and paraprofessionals who are on the school payroll and report to the principal, because the regular attendance of all these staff is important to student progress. Any staff member present less than 93 percent of the time (14 or more absences) is in the *below basic* category. This is not to say that some staff do not have legitimate reasons for an extended absence; only that the more staff are absent, the more student performance suffers.

The only part of the Performance Index that is calculated differently from the above examples is promotion and persistence. The promotion and persistence rates are not multiplied by a weighted value. That is because we do not have different students with different levels of performance on these measures -- there is a single rate for the school. The persistence rate measures the proportion of a school's ninth graders who graduate on time four years later. It gives a school credit for students who start at that school and transfer elsewhere in the system, and for students who transfer in after grade nine and graduate.

Once the Performance Index scores have been calculated for each indicator, the total score is calculated. First, the two enabling scores are averaged, to create a single enabling score. Then the reading, math, science, promotion or persistence, and enabling scores are averaged to create one total Performance Index score.

*Calculating the Performance Target:* The next step is to create the performance target. This number tells us the score a school should attain at the close of a two-year cycle in order to progress at a rate that will bring its average performance to the *proficient* level in 12 years. A school that has reached the 12-year goal has a Performance Index of 95. To find the total growth a school needs to make over twelve years, we subtract their baseline (1995-96) Performance Index score from 95.

The following chart shows how we would determine the total growth needed over 12 years for a school whose baseline Performance Index is 77.

12 year target score = 95

Baseline = 77

Total growth needed:  $95 - 77 = 18$

Over the 12 years (school-year 1995-96 through 2007-08), there will be six two-year performance cycles. Therefore, to set the growth target for the first two-year cycle, we take the total growth needed and divide by 6.

$18 / 6 = 3$

To continue the example, the school whose baseline Performance Index is 77 needs to reach a Performance Index score of 80 ( $77 + 3$ ) by the close of the first two-year cycle to be progressing at a rate that will bring the average performance to the *proficient* level in 12 years.

### ***Reduction in the Proportion of Students Scoring Below "Basic"***

There is a second part to the performance target. In order to qualify as meeting their target, schools must not only achieve the target Performance Index score, but must also reduce the overall proportion of students scoring below the basic level (including *untested* students) in reading, math, and science by at least 10 points during the first two-year cycle.

### ***Modifications Being Considered***

Among the modifications to Philadelphia's accountability system that are being considered or implemented are:

Assessing multiple grades per level: Beginning with the establishment of the baseline for Cycle Two (school years 1998-99 and 1999-2000), we will assess reading, mathematics, and science in two grades (3, 4, 7, 8, 10, and 11) for each level (elementary, middle, and high school). Including the results from two consecutive grades will help mitigate the effect on school scores of different cohorts of students. Also, assessing multiple grades per level for accountability purposes will promote a whole-school approach to instructional improvement and discourage schools from simply trying to focus resources on one tested grade.

Developing new assessments: We are in the process of developing new system-wide assessments. These assessments will supplement the existing testing program, provide additional accountability measures that are closely aligned with Philadelphia's curriculum standards, support the new promotion and graduation policy, and provide indicators of individual student progress that link to the accountability assessments.

Tagging students to sending schools: Currently, many students who leave their neighborhood school to attend special programs (e.g., some special education students and students who attend disciplinary programs) are counted for accountability purposes at the school where they receive their services. Beginning with Cycle Two, these students will count at their sending school. This change is designed to promote decisions about student placement that are based on the program most likely to best serve the student, rather than on which program is most convenient for the school.

Broadening measurement of persistence: We are considering an expansion of our definition of persistence at the high school level. Given the high dropout rate, low on-time graduation rate, and high failure rate in ninth grade, we may include a five-year persistence rate and a ninth grade promotion rate along with that of the four-year graduation rate. We want to reward schools for successfully engaging and educating their students.

Revising promotion and graduation requirements: The Board of Education is reviewing a proposal to revise the requirements for student promotion and graduation. Currently, promotion and graduation standards rest on report card marks and course credit accumulation. Under the revised policy, a variety of system-wide, performance-based assessment measures will be implemented in addition to requirements for marks and credits.

# Revisiting the Issue of Measuring and Reporting Student Growth in an Era of Standards-Based Reform

Sally J. Bennett  
San Diego City Schools

## Who are we?

San Diego City Schools is a large urban district with:

1. 137,250 Diverse Students: 34% Hispanic, 29% White, 17% African-American, 8% Filipino, 7% Indochinese. 27% English learners (LEP), 52 native languages. 14% gifted and talented; 10% special education; 60% poverty.
2. 167 Schools: 115 elementary, 22 middle/junior high, 16 comprehensive senior high, 14 multi-level and atypical.
3. 25,230 Employees: more than 8100 teachers.

## **Building a Standards-Based System**

Student achievement is the district's single most important goal and area of concern. To support the improvement of student achievement for all students, San Diego City Schools is building a standards-based system around five key elements. These elements, and the status of district work in these areas, include:

1. Content and Performance Standards to describe what students should know and be able to do, and at what level they should perform.
  1. District grade level standards in Language Arts and Mathematics approved by Board of Education in February-March 1998.
  2. District grade level standards currently being developed in Science, History-Social Science, and Visual and Performing Arts.
  3. Applied Learning Standards to be embedded into the content area standards.
2. Learning Environments (Curriculum and Instruction) to provide students with high-quality, rigorous content and focused, meaningful instructional programs.
  1. Work is beginning on aligning curriculum and instruction with the district standards, and building teacher capacity to implement standards-based instruction in the classroom.
3. A balanced Assessment system through which student progress toward and achievement of the standards can be measured and reported (to multiple internal and external audiences).
  1. Performance Assessment:
    1. Portfolios: districtwide literacy assessment portfolio being phased-in; reading exhibits will be scored districtwide at grades 3, 4, and 8 in 1998.
    2. Exhibitions: senior exhibitions are a graduation requirement beginning in 1998; will be included in district accountability system beginning in 2000.
    3. On-Demand Performance Assessment: will be added in 1999 or 2000 (may be the California Assessment of Applied Academic Skills, publisher-developed direct writing/open-ended math, or something similar).
  2. Norm-Referenced Testing:
    1. SAT 9 and Aprenda 2 for grades 2-11 (statewide tests beginning spring 98)
  3. Other Performance Indicators:
    1. Report Card Grades (standards-based report card under development)
    2. Advanced Course Completion (University of California a-f course requirements and career path course sequences)
    3. Voluntary Exams (including the California Golden State Exams, PSAT/SAT, Advanced Placement tests, etc.)

4. A clearly articulated, shared Accountability system which recognizes schools that are successful in improving student achievement and provides support to those schools that need assistance.
  1. District accountability system collaboratively developed by teachers, administrators, parents/community; approved by school board in April 1997.
  2. School is the unit of accountability; mutual accountability/shared responsibility among all stakeholders to improve student achievement.
  3. Long-range student achievement goals (90 percent of students meeting standards in ten years on performance-based assessments; 50 percent of students at or above the 50th percentile in six years on norm-referenced tests).
  4. Site-specific improvement targets set every two years from school baseline data on multiple student achievement indicators; targets set for all students and for ethnic groups demonstrating an achievement gap.
  5. Recognition for schools that meet improvement targets; support and intervention for those that do not meet targets.
5. A Public Reporting system through which student achievement information - individual student progress as well as school performance - is shared with teachers, schools, parents, community.
  1. Standards-based student report card in development (pilot in 98-99).
  2. Work underway on improving communication and public relations (internal and external) regarding student achievement/school performance.

-  
-

## Measuring and Reporting Student Performance

### The Challenges

### Where is the District Now?

#### Determining Student Performance

1. What does "meeting the standard" look like?  
How do we merge results from multiple indicators to determine student performance/progress?

1. Student work exemplars being collected for performance standards.  
Beginning work on merging results from multiple indicators to make student-level decisions.

#### Reporting to Parents

1. Has my child met the standards?  
What progress has s/he made?  
What does s/he need to work on?

1. Standards-based report card for K-8 in development; scheduled to pilot in 1998-99.

#### Using Data to Inform Curriculum/Instruction

1. What does this child need to

1. Beginning work on identifying and

enable him/her to meet the standards?  
What strategies does the teacher need to use/learn?  
What processes need to be in place at the school to support data-driven instructional decision making?

producing useful classroom-level data.  
Beginning work on building teacher and school capacity to use data to inform curriculum and instruction.

### The Challenges

1. What percent of students (all, disaggregated groups) have met standards? What progress has been made toward the district goals?  
How do we make the data/reports understandable and useful for planning and instructional decisionmaking?

1. How well are students in the district, schools meeting standards? What progress has been made?  
Which schools should be recognized? Which need support?  
How do we make our public reporting clear and understandable?

### Where is the District Now?

#### Reporting to Schools

1. Accountability data sheets provide: baseline data and two-year targets; percent of students meeting standards; schoolwide and disaggregated data.  
Summary sheet with narrative interpretive comments.  
Continuous self-study model being implemented through accountability system and other review processes.

#### Reporting to the Public

1. Accountability system includes process for identifying schools for recognition and for review/intervention.  
Working on public relations and communication plan to better report school performance results.

## Continued Challenges

- Finding the right combination and balance of indicators to best measure student and school performance.
- Addressing the expectation to measure every student - every standard - every year.
- Appropriately assessing special student populations (e.g., English language learners, special education students).
- Appropriately reporting and utilizing disaggregated data.

- Balancing absolute performance with progress.
- Balancing fairness (having a psychometrically and statistically sound system) with understandability (having a system can be easily and clearly communicated to various publics).
- Dealing with increasing demands for data/results from multiple audiences.
- Dealing with technical, implementation, and management issues.
- Getting changes into the classroom.
- Making systemic changes.

# Discussant's Comments

H.D. Hoover

University of Iowa

I want to say something here that at the beginning. I only saw one of these papers, and only part of it, before today. And because of that it's probably going to appear that I'm unfairly picking on one of them. So I guess maybe the lesson would be that in the future don't ever send in a damn paper.

You obviously know this Mitch, but in fact it is closely related to what I'm going to focus on when I look at the Philadelphia data, and it's very closely related to the way things were done in Kentucky. There is absolutely nothing I'm going to say that has anything specifically where I'm trying to be nasty to Philadelphia. Joanne [Lemke, Harcourt Brace Educational Measurement], I'll point out some points here but that sure isn't jumping on the Stanford 9. That's for sure and you'll see what I mean in a second why that is so. But I will make a couple of comments on the other papers.

I would make sort of an opening statement that nobody talked about growth. Period. I mean some people could, you know. Ed, used the word a little, but he was not really talking about growth very often and the measurement of growth. So if you came to this session expecting to learn something about growth, I hope you don't leave thinking you have.

[Laughter] Which might say something about how much you knew about it when you came in. [Laughter.] But, I'd better leave that alone. Now, I will say this, I got two things from Philadelphia – Children Achieving Action Design Early and then I got the Professional Responsibility Index. Which it ends up with something that I'm sure is not quite the same as what you put up on the overhead but I was amazed at something here, and since many of you know what I know you know why I bring this up. Children Achieving Action Design, which as the 95-99 schedule says, "Implement a system," now this for Philadelphia remember, "of performance based assessments tied to the new standards for students," and then it says, "beginning in Spring 1995 we will replace norm-reference multiple tests which tells us nothing about whether the students can meet absolute standards. With performance based assessments which ask students to demonstrate their mastery of academic subjects in real world tasks. These new assessments will be sensitive to our students diversified backgrounds."

JoAnne, is the Stanford still a norm-referenced standardized test? [Laughter]. I thought it was, because it has kicked the ITBS's butt in lots of situations, and I thought we were sort of doing the same kinds of things. And, in fact when they sent out the answers on the professional responsibility index, the answers that they sent to teachers addressed this. Obviously some other people brought this up and said how is the SAT9 different from previous system wide measures. And it says "the SAT9 is much more performance-based. It's also based on higher standards and derived from the national standards on which the Philadelphia standards are based."

I wanted to make it very clear here that the Iowa Tests of Basic Skills is much more performance-based. [Laughter]  
Now, I hope I haven't lost my overhead, because this is, I have made this little sucker up, and those of you who know me know I like to do these things. I always say this is an H.D. Hoover PowerPoint presentation. I always end up talking about numbers and people think that I am this weirdo mathematician type, you know, but in fact this is pretty important stuff. This is the, now my numbers here are a little bit – maybe this was the primary level or something. I don't know, this number, this is one in the paper I got. This is reading, the one you put up there was 51.4, which I assume was a different grade, or maybe it was a combined elementary and secondary. It doesn't make any difference to the point. In fact, those numbers that Mitch put up there were more demanding because that index that he got is 51.4.

Now, remember how these are arrived at. It is important that you see here – you need to pay some attention to this column here about the percent of students. This is very important. Think about what this means. In Philadelphia, in this set of data anyway, 6.4 percent of the kids were advanced, 8.7 percent

were proficient, and 18.9 were basic. This is baseline data. Now what that really means is that 66 percent of the kids were below basic in this context, which I don't know where ever that cut point is between here and here. That might be somewhere like around the national means, and I see maybe 66 percent of – probably in that vicinity anyway. I don't know how close, but it's close enough I think it's something we really want to think about what this means. And what it means for people to go out and set standards. By the way, this is the first thing though that I've seen where people have used the word standards and they set one. We talk about it a lot, but we don't have many.

These are standards and they are standards to be met. Now, the goal here is that Philadelphia will have, in 12 years, an average on this of 95. Now it ends up that there are five different pieces that go into those. But the one that I've chosen here, reading, is the one where they can most easily meet it in the three academic ones. The other two they can get up into the 90's much easier, but still there, they are still going to be somewhere in the 90's when they do this on the other two.

Now, what has to happen for Philadelphia to meet goal? Now remember Mitch indicated that 80 percent of the kids in Philadelphia are receiving free and reduced lunch. Right? That's the indicator there. For that to happen, and this doesn't even quite make it. If after these years, 30 percent of the kids in Philadelphia are advanced, 30 percent are proficient, and 30 percent are basic. And let's take these ten percent down here that they are not tested. And, by the way, this is a pretty important point about this improvement and the points Mitch made. Now I think this is what, by the way you will notice that these are actually below basics one two and three. This is related to what Ed [Ready] was saying. This was in the paper, the part I got, broke those down further because everybody was lumped into this one place. Now, and those actually get different points. Now if you don't test, you don't get any points. Now the first thing I'd do if I had this system, after the first year, and I had ten or twenty percent of the kids not test. I'd say test the little suckers. [Laughter.] I'd get some points.

I can sure show some serious immediate growth. Now it may not persist and I'll tell you all what you only want to do that the year that you think that this is the year that I gonna get my ass fired. Save it for when you need it.

Even if we do this, using this index, and I can get 30 percent of the kids advanced, 30 percent proficient, and 30 percent basic. And then I take those ten percent of the kids I'm not testing, because I guess I just assumed they can't do anything or something, I don't know what all the reasons are, and we know there are many. But at least I test them. Well, immediately I'm going to get two points out of them that way. I get ten, there's ten multiplied by point two

so I get two. By the way I still don't have this up to 92. I don't have 95 yet which is what I have to get.

And, there is another thing that is very important to look at this scale and ones like it. When it is set this way, and I have no arguments with advanced-proficient-basic in any real sense. But when you have said it this way and some population and for the sake of this point I'm trying to make, is I think that cut score to get above basic appears to be in the vicinity of the national median, the 50<sup>th</sup> percentile.

What this says is that Philadelphia, in the next 12 years, with 80 percent of their kids receiving free and reduced lunch, has to get 90 percent of their kids above the national average. This is silliness. I mean, it is. You have to look at what this kind of scale says and what it means. And I am absolutely not against higher standards. In fact it is wonderful that we have higher standards, but to give you a little bit of a feeling for this, I actually, the I reason I get to do this is I actually do run state testing. Well that's not true. I have something to do with the state testing program. Not a lot, but a little – the one in Iowa.

Now people happen to know that Iowa actually scores pretty good, and frankly recently when they showed on the NAEP and Iowa was the state that looked like it had world class standards because we scored as high as Korea. That's true, that's pretty awesome. Now I want you to look, this is the data for the state of Iowa compared to national norms. And by the way there is nothing sneaky here, this happens to be on the ITBS. But it looks exactly the same way on the National Assessment of Educational Progress. Which shouldn't surprise anybody if they are both norm-referenced tests, which they sort of are, and are supposedly based on representative national samples. But look at this – in Iowa we don't have 80 percent of our kids in free and reduced lunch. But you know what, we only have an average of 66 percent of the kids in the state of Iowa above the national mean. And we are the highest scoring state, well we are one of about five that sort of trades places as the highest scoring state in the United States. But, we are never going to have 90 percent of the kids in Iowa score above the national average unless we don't test half of them. [Laughter.]

Actually, you know where I'm from. We're going to let Missouri annex half of the state. The lower half.

In no sense am I trying to pick on Philadelphia, I'm really not. Now I do think that this scale it is sort of surprising that these weights that we are talking about in Kentucky, and maybe I'm wrong about this but the ones in Philadelphia, sort of came from the same place. That the ideas behind using this kind of weighting system came from the same place, but I may be wrong. That's why I won't say anything more than that. But this is pretty important that we think about this when we set standards. And there is nothing wrong with setting them, that we

set ones that in fact are in some sense are attainable and are reasonable. Now Ed said that we should have looked at "is this reasonable?" Well, I think you could have looked at this and said "that doesn't seem very reasonable." Because I don't think it is. Now I've made 17 enemies, and that's the short list. [Laughter.]

By the way, let me show you something else. I always like to show this because I actually want to make another point. This is my Ronald Reagan graph. That's the achievement for the state of Iowa the last 40 years. Grade eight, grade seven, grade six, they went up and down. Remember Ronald Reagan was the great communicator. He was the greatest I've ever seen presenting data. And this is why I call this my Ronald Reagan graph, because back here, you see that, one up, that's good, bad, good. [Laughter.]

Ronald Reagan would give these talks and, I'm a statistics professor and I've learned more from Ronald Reagan than anybody I know. He would give these talks, and he's going along and you'll remember him, he was a cool dude. He was rolling along on these talks and he decides to trash the Democrats, which of course was always half of one of the talks. So to trash the Democrats what he would do was say something nasty then he'd whip up a graph. Always it would be color stuff, I mean, it would be fancy. But always to trash the Democrats, he would show something like this. That suckers going that way down. Now it didn't have any numbers on the scales or anything, we don't need those. But it was going down hill to beat the band. Now, actually the reason I want to show this is for a couple of reasons. One is I'm afraid in Iowa if we switch to standard based measurement we'll lose all this. But, first of all this is not a measurement of growth. This is status across time and how it has changed. And, it simply shows how the system has held from year to year, getting better, getting worse. But the important thing to show here is, you may notice that in Iowa the last few years the achievement is going down a little. And you might also note that nationally the latest NAEP data, and they are being really quiet about it, it's going down a little. And if you think that this is just Iowa it isn't -- I have one for the nation that looks just like it.

The same thing makes test scores go up and down everywhere, and I sort of think we are getting ready to have a time when test scores are going to go down a little bit. Now that's going to make it harder for Philadelphia. Not easier. It's going to make it harder for everybody. And so, everything we are hearing is assuming that we have all of these standards and obviously everybody is going to keep going up. Well most evidence that I see over the last few years indicates that is not the case. That's especially nerve wracking or frightening or whatever to see that achievement in the lower elementary grades for the first time in the state of Iowa was declining. First time ever. And, we don't just have the caucuses folks, you know -- first in the election, first in achievement, first up,

first down. And back there in the sixties we bottomed out before anybody else I think,. So that is something that is related to all of these issues . When you establish these standards and you have to keep in mind that behind all of this you have things happening nationally that may or may not be causing achievement to go up or go down. And those things, of course, are incredibly complex.

Now, oh my gosh I bet these other authors are hurt that I lost my notes. [Laughter]. I did have this one great line, though. I'll go ahead and use it. San Diego said that there, this seems to be pretty reasonable, they want half of the kids to be above average in San Diego. But you want 90 percent of them to meet the standards. Well Philadelphia wants 90 percent of them to be above average and 95 percent to meet standards. Now that's something about the non-linearity of those transformations that you are talking about there. We got a lot happening right there in that five percent.

In the case of San Diego I am anxious to see the very data that you are talking about. I mean about what's happening on these performance assessments, I could ask you questions about how are you going to keep those scores comparable in some sense from year to year. But I do hope that we do start seeing more data from things like this, and Mitch had the misfortune of showing me some.

Again I want to point out that what they are talking about in San Diego is status not growth -- status this year, status next year, status the next year. We are not talking about measuring kids and saying okay does this kid this year know more than they did last year? I might have some disagreements with Ed regarding the quality of the developmental scales that are associated with certain kinds of tests. Frankly, they're not perfect but they are the best we got. And, because if there was something better than we'd have all of these smart people that would have figured them out. And in fact they do assess growth and you can even have norms for growth if you want them. And, whoops, I said that word again, I know that's something I'm not supposed to say at this meeting.

I think one of the great things that Milwaukee is doing, is the involvement of the teachers, the staff development nature of it. I wonder, though, what is growth in those situations. You said student achievement, but how do you know? Because Ed doesn't think that we can measure it, and I don't see any evidence that you are. So if we say achievement is getting better or that we are getting growth, I don't quite know where that is coming from.

Ed, what does proficient look like? I always wonder about these things, that we're all at once gonna help people by showing them what proficient looks like, or what advanced looks like. The minute that you show me a test and show me

a kid's score and say that is what proficient looks like I'll show you another kid who has either the same score or in which proficiency for that kid won't look anything like that, or another child who we might say the same thing about who is not near as proficient. I think this idea that we are describing very accurately for people what proficient is, what advanced is is not quite accurate. We still have a long way to go.

# Reply To Discussant

**Mitchell D. Chester**

I am both grateful and disappointed with H. D. Hoover's critique of my presentation today. I am grateful that Dr. Hoover was frank about his skepticism that poor students can learn at high levels -- this is an issue that is on the minds of many, and one that is too often left unstated. I am disappointed, however, with the substance of his critique. There are a number of issues that H. D. could have raised that would have fostered a substantive discussion of standards-driven accountability systems. Constructive discussion of the few accountability systems that have been implemented is needed, since the enterprise of measuring school and district growth is still in its infancy and we have a lot to learn. Unfortunately, Dr. Hoover's remarks this afternoon did little to further this discussion.

The undoing of H. D. Hoover's discussion was his use of a norm-referenced paradigm to both (1) attack the credibility of a system that expects poor, historically low-performing students to achieve at substantially higher levels and (2) critique a standards-based measurement system. H. D.'s argument essentially distilled to this:

Philadelphia serves primarily poor students [more than 80 percent of our 215,000 students qualify for free or reduced lunch]. Philadelphia's goal is to have at least 95 percent of their students demonstrating *proficient* reading, mathematics, and science achievement within 12 years [by the year 2008]. Currently, fewer than one-half of Philadelphia's students are reaching even basic achievement levels on standardized assessments in most grades and subjects. At the same time, relatively few students nationally are demonstrating *proficient* achievement levels. Therefore, according to Dr. Hoover, what Philadelphia expects is that their students, who are mostly poor and rank near the bottom of the norm-referenced distribution, will leap-frog most of the nation's students and perform at a percentile rank that is above the national average.

Dr. Hoover's critique is a striking example of the limitations of norm-referenced testing, and the inappropriateness of using a norm-referenced paradigm to critique a standards-driven accountability system. The norm-referenced paradigm is built on sorting and ranking, and developed to determine where a student lies in comparison to a group of peers. The objective of norm-referenced testing is to spread students out in the bell-curved distribution. Standards-based assessment systems are designed to measure achievement relative to expectations for student performance (performance standards) in valued content areas (content standards). The objective of standards-based assessment is to determine what students know and can do. The standards-based perspective is interested in whether students have demonstrated levels of understanding and competence with specific subject matter that are needed for success in the world and, if they have not, the level of understanding and competence they have realized.

Standards-based assessment instruments are designed to accurately and sufficiently measure student performance on those dimensions of the curriculum that our reforms and improvement efforts are designed to effect. Over time, results from standards-based assessments will provide an indication of students' progress relative to the curriculum if they are benchmarked to performance standards that are held constant. Even though student achievement may be improving in a school or district, there are at least two reasons that norm-referenced testing results may not reflect this growth. First, the norm-referenced test may have little fidelity in relation to the curriculum standards that are the focus of the reform. Second, if the attainment of the general population is improving over time, then the achievement of any one district or school may not be reflected in percentile ranks. If everyone is improving at roughly the same rate, for example, then any given institution is likely to maintain its ranking relative to successive norming samples.

As disappointed as I was with H. D. Hoover's inappropriate critique, I was equally disappointed with the fact that he failed to raise any of the many substantive

questions that are being asked about standards-based accountability systems. Among the issues that are being raised are a variety of questions related to standard setting, measurement, and inferences about growth. For example: What are the theories of institutional growth that underlie accountability systems, and are they adequate for establishing progress that can reasonably be expected of a school or district over a given amount of time? How can research on teaching and learning help us to design accountability formulas that set fair and reasonable achievement targets without lowering expectations for poor students? How do we establish cutpoints that minimize errors in the classification of schools? These and many other questions could have been raised by H. D., and would have fostered constructive inquiry about the topic at hand -- as a presenter, I would have welcomed such discussion.

Dr. Hoover's comments reinforce the beliefs of many who wonder whether anyone can substantially improve learning in city schools. While many urban educators accept responsibility for the fact that most urban districts perform poorly on almost every academic measure and have implemented reforms to improve their students' achievement, a growing number of policymakers and politicians have lost faith in the ability of those educators to promote better learning. Sometimes that lack of faith is based on the belief that poor and minority students cannot learn to high levels, that urban educators cannot deliver substantially better performance, and/or that further fiscal support of urban districts is a waste of taxpayer dollars.

While I must admit that my initial reaction to H. D.'s critique was one of disappointment and concern, I believe that he should be given credit for raising this perspective -- too often it is the subtext and not the storyline. Unfortunately, Dr. Hoover's critique plays into the hands of those who believe that educators who serve poor children cannot make a difference in the achievement of their students. If we believe that schools *can* make a difference, and that student learning *can* be influenced by changes in curriculum, instruction, professional development, and programs, then it is plausible that district reform and school improvement initiatives *can* impact student achievement.

I am most disappointed, however, that an accomplished psychometrician is apparently unwilling to consider the world from other than a norm-referenced paradigm. Dr. Hoover's discussion this afternoon was both inappropriate in its juxtaposition of norm and standards-based paradigms, and bereft of substance despite the rich possibilities for discussion.