

NATIONAL ACADEMY OF SCIENCES  
NATIONAL RESEARCH COUNCIL  
BOARD ON TESTING AND ASSESSMENT

SYMPOSIUM ON THE USE OF  
SCHOOL-LEVEL DATA FOR  
EVALUATING FEDERAL  
EDUCATION PROGRAMS

December 9, 2005

U.S. Department of Education  
Barnard Auditorium  
Washington, D.C.

Proceedings By:

CASET Associates, Ltd.  
10201 Lee Highway, Suite 180  
Fairfax, Virginia 22030  
(703) 352-0091

*Copyright © 2005 National Academy of Sciences. All rights reserved. No part of these pages, either text or image may be used for any purpose other than personal use. Reproduction, modification, storage in a retrieval system or retransmission, in any form or by any means, electronic, mechanical or otherwise, for reasons other than personal use, is strictly prohibited without prior written permission.*

*Opinions and statements included in the paper are solely those of the individual author(s), and are not necessarily adopted or endorsed or verified as accurate by the Planning Committee for the Symposium on the Use of School-Level Data in Evaluating Federal Education Programs or the National Academy of Sciences, including the National Academy of Engineering, Institute of Medicine, or National Research Council.*

List of Participants:

Steve Dunbar

Robert Linn

David Thissen

Donald Rubin

Sam Lucas

Gage Kingsbury

Kashka Kubzdela

Geno Flores

Steve Henry

Robin Taylor

Joe O'Reilly

Cory Curl

Mitch Chester

Wendy Yen

Don McLaughlin

Laurie Wise

David Francis

Diana Pullin

William Schafer

## TABLE OF CONTENTS

	<u>Page</u>
Session 3: Cross-State Comparisons	
Robert Linn	1
Discussion Panel	
David Thissen	24
Donald Rubin	36
Session 4: New Opportunities	
Gage Kingsbury	58
Kashka Kubzdela	75
Session 5: State and District Reactions	
Robin Taylor	84
Joe O'Reilly	85
Cory Curl	85
Mitch Chester	86
Session 6: Larger Contexts for the Database	
Don McLaughlin	130
Session 7: Synthesis Discussion Panel: Balancing the Policy and Technical Issues	
David Francis	155
Diana Pullin	162
William Schafer	171
Laurie Wise	171
Closing Comments: Steve Dunbar	194



P R O C E E D I N G S

**Agenda Item: Session 3: Cross-State Comparisons**

DR. DUNBAR: I thought we might just allow a little bit of time immediately after Bob's presentation for any clarification questions, and then we will move directly on to the discussions.

DR. LINN: Good morning, everybody. Was anybody surprised that there was a two-hour delay this morning?

Well, it is nice to be back here. I didn't know how many people would be here after yesterday, with the weather coming up here. Let me get started.

What I was supposed to prepare a paper on was looking at the possibility of adjusting for differences between tests, so that you can make comparisons across states, as the primary interest.

In doing that, one of the things about getting old is, you start remembering when you were in graduate school and things like that. It turns out that this desire for being able to adjust for differences to equate tests if you will, or link them together so that they can be interchangeable, is not exactly a new idea.

When I was busily working on my dissertation in 1964, there was a symposium at NCME, National Council on Measurement in Education. It had four mid-20th century superstars, John Flanagan, Bill Engoff, E.F. Lindquist and

Roger Lennon. In that symposium, they were looking at equating non-parallel tests, the topic of the four lead articles in the first issue of the Journal of Educational Measurement in 1964. Not much publication way back then, I guess.

So anyway, I had a quote from each of the four of them. You can read them as well as I can. Flanagan was pointing out something that was on peoples' minds later when the NRC committee that I will talk about looked at the issue of trying to link all the tests together across the country. It would be group specific. You can always get a conversion, but the conversion you would get on one group would not hold for another group.

Lindquist similarly was arguing that if you have non-parallel tests, you just can't use them interchangeably. And Roger Lennon made the point that what you have in a conversion is going to be group specific.

Now, these reservations of these four individuals that had very high status within the field of educational measurement at that time didn't dissuade anyone, as far as I can tell. Right after they published their papers, the ESEA was passed, the first version of what is now No Child Left Behind, through several reauthorizations. That was passed in 1965, and that led to a desire to be able to aggregate scores across states.

Now, it actually was a similar problem in those days, because most everyone was using one of six or eight different standardized tests, and so you weren't talking about 50 different tests. Not that every state has an entirely different test, but we have moved in that direction, certainly.

So this fed this desire to be able to aggregate results across states, and that led to the thinking about, could we do a big equating study. The anchor test study was -- a lot of the design for that was in the mind of Dick Yaeger at the time he was working for the Department of Education.

So they brought together some committees and looked at the different standardized tests, reading tests, math tests. After reviewing the content of the math test, they concluded that it was not reasonable to try to equate all the different math tests. They just were too different in terms of how much emphasis was put on basic skills, how much was on one kind of mathematics versus another. But they decided that in reading, there was enough commonality that it made sense, at least at the elementary grade levels, to try to equate them.

So this very large national anchor test study was launched. It had looked at grades four, five and six in reading only. The study I would say was largely

successful. You had the eight major reading tests at those grades that were put on a common basis, linked together. They did a lot of checking, something that is an idea that I think more equating studies ought to be looking at. That is, the degree to which the transformation you have is really invariant across subpopulations.

They looked at that for various subpopulations in the anchor test study, and the results were pretty promising. They weren't perfectly invariant, obviously. Even versions of the SAT would not be perfectly invariant, but they were reasonably close, and so that provided some assurance.

The other thing that the study did was take one of these tests, the metropolitan reading test, and say that they would develop norms on it, national norms. The national norms were better than any of the individual test publishers were able to do because of the nature of the cooperation that they got that is harder for individual test publishers.

So it was pretty successful. On the other hand, it was limited to only one subject, reading, in only three grade levels. Furthermore, it wasn't long after the results were out, and you could start making a conversion from one test to another, that every test publisher that was included in the study revised their tests. So you had

a new set of tests out there that were being administered for Title 1 purposes.

So Title 1 went in a little different direction. They said, what we will do is, we will just let the publishers' percentile rank carry the day, only they converted them to this, at that time a new scale that no one had ever heard of called normal curve equivalents [NCEs]. Certainly we had heard of normal transformations and standard scores, but the NCEs were thought to be pretty clever, because they corresponded with the percentile ranks at three points, so they just set the standard deviation, the mean, to 50 so that would correspond to a percentile rank of 50, and then a standard deviation so that a score of one or 99 on an NCE would be the same as the percentile rank.

So this was a kind of standard score aggregation, and results were then aggregated within states, because districts were often the ones deciding which of the standardized tests were going to be used, and then aggregated it again up to the national level.

The assumption that it is okay to do that was never, so far as I could discover, seriously challenged. I think that is largely because nobody paid much attention to the aggregate results when they got aggregated up to the federal level, not even at the state level much.

Well, if you go forward a bit to the standards-based assessment movements, which is continuing today, all of a sudden the challenges become a lot greater, because each state by requirements in the forerunner of No Child Left Behind, IASA, back in the '90s when that was a reauthorization of Title 1, encouraged the states to develop their own content standards and then have assessments that were in line with them. Now, IASA did not have the kind of teeth that No Child Left Behind has, but in any event, there was a movement in that direction.

Well, Clinton was President. In the State of the Union address in 1997, he called for something that upset both the left and the right. The left, because as Chuck Rebenon has quipped, the left was upset because he was calling for a voluntary national test, and the right was upset because he was calling for something national that would lead to a national curriculum.

So as often happens, when Congress got upset about this notion of a voluntary national test, they came to the National Academy. The National Research Council put together several different studies, one of them to look at the voluntary national test itself, and another one to ask the question, if we have all these tests, wouldn't there be some way to find an equating. So it is back to the 1964 question. These tests are not really parallel, but can't

we equate them in some way.

So this led to the charge to the one committee that led to what is known as the uncommon measures report that was chaired by Paul Holland. This was the charge to that committee, that they were supposed to investigate the feasibility of developing a common scale that can link all the existing commercial tests and tests that the states might have in one form or another so that you could move from one to the next and treat them interchangeably.

Those of you that know Paul Holland, and Fred Mosteller was also on the committee, these are statisticians that are consistent with Judy Singer's comments yesterday, they don't want to be presented with a problem and then come back with an answer that says no, you can't do this. So they struggled with it quite a bit, because their view is that the statistician should be providing help, not just saying you can't do something.

But this is the conclusion that the committee ended up with, the first and primary conclusion, that it just wasn't feasible to compare this full array of tests through the development of a single equivalency linking scale.

Now, why is that? There has been a fair consensus, I would say, in the field of educational measurement in looking at what is required to be able to

equate. A fairly recent paper by Neil Dorans and Paul Holland lays out the five conditions or criteria for figuring out whether or not you can equate tests. These are listed here. You have the same construct, and that you have the same reliability -- reliability is important because if you are a person below average you prefer a less reliable test because it gives you a better chance of getting closer to the mean, but if you are somebody that is way above average, you would like something that was quite reliable -- the equating function should be symmetrical so that you go from X to Y or Y to X, you shouldn't have different functions doing that as you would if you were doing two different regressions, and there should be equity so that it would be a matter of indifference whether they got form X or form Y. Then you have this population invariance, the thing that was looked at in the anchor test study.

There is a fair amount of research that attempted to do things like equate education tests to NAEP. When that was looked at more carefully, the general conclusion was that the linkages just don't hold up in a way that you would need them for an equating. This is from a paper that Dave Thissen gave at ETS this summer, in July, which I think sums up well the reason that you can't have the invariance that you hope you would have on the equating.

The reasons for that are simply that if the states have different content standards, this implies that they also are measuring different constructs, or trying to measure different constructs with the tests that they put together based on those content standards.

If you look around the states, they have different numbers and different types of items. Some have more constructive response, some have multiple choice. They are of different length, and if you have different length you can be quite sure you are going to have different reliability.

If you think about what the content standards in the state tests are about, it is to encourage instruction to move in a particular direction. So to the extent that that happens within a state, you are going to fail to meet the equity requirement, because if you are targeting a particular set of content standards or teaching it in the case of the teacher, it is going to matter to you whether you get that state test or some other test that you weren't studying for.

As I said, there was discomfort among committee members of the common measures report, in part because they as I said would like to come up with an answer, that you could do certain things, even if you couldn't do everything that you wanted to do. There was also a feeling that there

was a lack of any good quantitative index of how bad it would be if you treated scores as if they had been equated or linked in some way.

So there was another conclusion in the uncommon measures report which I will just let you read, that tries to qualify things, saying that it might be possible to calculate this linkage, but that it would depend upon a bunch of factors, like how similar the content was, how the tests were used, and to the extent that they differed on these factors, it suggested that the linkage might not be that useful.

Now, one problem with that is that it was not very clear about which of these factors is really going to matter, and what interpretations are going to be justified, as opposed to which ones are not, so what the criteria are for doing that.

Well, Congress didn't like having an answer of no, so there was another committee started. This is known as the embedding committee, and it was chaired by Dan Koretz. The idea of embedding was, instead of trying to equate all these tests or link them together externally to NAEP, what if you took some blocks of NAEP items and included them in the state tests, embedded them in the state tests.

This committee also came to the conclusion that

was consistent with the uncommon measures report, saying that while you could physically do that, it was not going to lead you to a common scale that would have the properties that were desired.

Now, I think a problem, not just with these reports, but with the field in general, is being able to sort out what inferences can be validly made if you make a linkage, and which ones cannot. For that, you need more in the way of a quantitative index that tells you how far short of a satisfactory equating a particular linkage is.

That is a problem that Neil Dorans and Paul Holland in particular have worked on, and several other people too, looking at how can we index the degree to which you really lack meeting this criteria of invariance across different populations. So they used a couple of quantitative indices to track the degree of invariance that you have. They applied these indices to a bunch of different situations, some of which everyone would agree that the equating is just fine, two forms of the verbal SAT, one to another, for example, or two forms of the ACT mathematics, one with the other. Then they also applied it to situations that everyone would agree is a silly thing to try to equate, like the verbal and math tests. So you can do the arithmetic for a verbal and math test, but if you do it, what you expect to find is that the invariance is going

to be very poor.

That is exactly what these indices show. They also provide some intermediate information, like situations where people think it is reasonable to at least try to have some kind of conversion, and that would be as an example the SAT math with the ACT math. These are obviously not parallel measures, but they are at least targeted at the same population of potential college applicants, and they are both mathematics tests.

What they found is that indeed, that sort of concordance linking has an intermediate value of these indices. I think that is encouraging as a way of trying to sort out and quantify the degree to which you are falling short of this goal of having tests that are equated.

NCLB has just increased the demand for comparing the results. The comparison happens with or without us, so to speak. If you look at the recent article in the New York Times, for example, comparing the proportion of kids meeting standards on state assessments versus meeting it on NAEP, I have been guilty of doing those kind of comparisons myself as I will show you in a minute. NCLB also has this new requirement. It used to be that most states participated in state NAEP, like 40 or so, before they had this requirement, but in NCLB all states have to participate in reading and math at grades four and eight

every other year.

We just had the second round of that, the first being in 2003 and the second in 2005, where every state is included in the NAEP reading and math at grades four and eight. So that provides you with a better basis of looking at the possibility of linking state assessments to NAEP.

Why that may be of interest is illustrated by this graph of results for 33 of the states that had mathematics assessment results in grade eight in time to be included in Lynn Ulpton's Ed Week piece. This shows the percent of the students who are proficient or above at grade eight according to the state assessments.

You see some peculiar things. One is obviously that the variability is huge, and furthermore, it doesn't seem to make a whole lot of sense. For example, you have a comparison here of Missouri and Tennessee, where the proportion meeting the standard in the state jumps from 16 percent to 87 percent. It seems intuitively not very possible that the mathematics treatment in grade eight in Tennessee is that remarkably much better than the mathematics treatment in Missouri.

This doesn't necessarily imply that the test is harder in Missouri, but it does show that the standard, or some combination of where the standard is set and the test difficulty. If you look, on the other hand, at state NAEP

for these same 33 states, you see the variability is much less. Furthermore, the variability that you do see, which states are high and which states are low, makes more sense in terms of what other things you know about those states, like performance on the other tests, what the education level is in the state and so on. So the high percentages in the state NAEP results at grade eight and a half are Connecticut, Montana, South Dakota and Wisconsin, and that compares to the low percentages of Alabama, Mississippi and New Mexico, which I think arguably makes a whole lot more sense than the differences that you saw on the state assessment.

If you plot these one against the other, what you also note is that there are only two states that are above that line of agreement in terms of percent proficient or above. Most of the states are below it, all but two obviously; that is 31 out of 33. The relationship is pretty weak. Remembering that these are state aggregates, the correlation is only .34. I compared that to what the correlation would be for fourth grade reading with eighth grade math, and that correlation is much higher than .34. It is in the paper. I didn't put it down here what the number was. So now we are talking about mathematics tests -- two different mathematics tests at the same grade as opposed to different content, different grades.

Well, this whole symposium is about this database, so what does all this have to do with the database? As we talked about a lot yesterday, the hope is that it can be used to make causal inferences about federal programs. We talked a lot about the challenges of doing that. My focus is not on that obviously, but rather on using the database to make these linkages between state assessments and NAEP.

Don McLaughlin and his colleagues have focused on this quite a bit over the last several years. In 2002 they gave a paper in which they looked at fourth grade mathematics using the 2000 assessment for 29 states, where they had data.

This is a chart from that report showing what happens when the performance standards on the state assessments were linked to NAEP. What you see is, that red line is showing situations where the precision is fairly good, and the lower level standards in particular often are not so good. But for the proficient and advanced levels, there seems to be a pretty good level of precision.

What you saw in that chart is that the state standards span over 100 points on the NAEP scale. The NAEP scale is set up so that within a grade level you have a standard deviation of something like 40, so that 100 points is a pretty wide spread for looking at something like

performance standards that might have the same name, proficient.

They have also looked at state reading assessments at grades four and eight, and found results fairly similar to what was just shown there. They are not alone in doing this. Henry Brown at ETS has done some work, a kind of refinement of Don's technique, that takes into account using the NAEP sampling weights, injecting NAEP procedures, to come up with what Henry calls a weighted aggregate mapping to get the estimates of what the linkage is between the state assessments and NAEP.

They looked at the 2000 mathematics assessment results again, the same ones I just showed you, that Don and Victor [Bandeira de Mello] had done, and looked at it using their technique. What they found is that the mapping looks pretty much the same. There are very similar mappings with the two techniques. They also show some fairly substantial differences. Here is an example of a 95 percent confidence interval for two states that do not come close to overlapping, and where the proficient level is set in those states.

Well, it seems to me that it is clear from those examples that the database can be used to make the linking, to get some sense of where the state performance standards fall. It does have a good deal of uncertainty, so this is

not as precise as you would want when you are equating and you start using those scores interchangeably, but at least it gives you ballpark figures. Henry Brown and his colleagues have suggested that they don't think this approach should be used for making a big deal about small differences, but they suggest that with confidence intervals you can at least see conversion differences and have a good judgment on that.

It is not that clear to me whether or not the advantages and disadvantages of using the linking like this as compared to something else, like looking at standard scores, or back to the good old days of NCEs, the norm equivalents, to be able to have your aggregate results to make some kinds of comparisons.

An attempt at a few conclusions here. Number one, I don't think this desire and request for coming up with common scale is going to go away. If anything, it seems to me it has just gotten stronger over the years.

The second thing is that NAEP can be used to link state standards to the NAEP scale. Don and Henry's work, for example, illustrate that. There are, however, considerations -- you have to think about how this is going to be used, because the stability of the links to NAEP and the interpretations may be questionable. The linkage you have in 2003 may not be the same as the linkage you have in

2005, so that is a lack of invariance that is important if you are going to be tracking things over time.

People who object to using NAEP in this way usually point to the different contents in the motivation as well as to the differences in the content. The framework for NAEP is not the same as I said yesterday as the content standards of various states.

Dave Thissen put it this way last summer at the ETS linking conference. It is a good rhetorical question, which summed up very well the notion that motivation may in fact be a critical element if we are going to try to use NAEP in this way.

The fifth thing is that it does seem clear to me that the strict equating, the requirements for it, the ones that I listed from Neil Dorans and Paul Holland, are not going to be met for the linkages that people are interested in between state tests and NAEP, or between one state test and another.

Finally, I would like to see a good deal more work along the lines of Neil [Dorans] and Paul Holland to have a better way of judging just how far away from achieving invariance that you want for an equating we are with different kinds of linkages.

I'll stop there. I can stay here if there are questions or come down there, either one.

DR. DUNBAR: Are there questions for Bob, clarifications on points in the paper?

PARTICIPANT: I have two questions. One, when you mentioned that the stability of the links is possibly questionable, is it possible to look at that a slightly different way? If you make a link say in 2003 and then you go back and you make a link in 2005 and you find that they are different, is that going to be a basis for judging whether state standards have changed?

DR. LINN: Well, that would be one interpretation, but it could also mean that the difference in how much work has gone into preparing kids for the state assessment over that two-year period may have moved while not moving performance on NAEP very much. So it would be ambiguous as to which of those would be true.

I did a study a number of years ago, in which we linked state assessment to NAEP, and looked at the stability over -- I think it was a two-year period, I can't remember for sure, well, it would have to be, because NAEP is only administered every two years -- looked at the stability, and the stability was not very good. So the equating function that you had in one year was not the same one that you would have two years later.

So that is a kind of invariance that I think we would like to have if we are going to treat these scores

interchangeably.

PARTICIPANT: The second question is maybe off the wall, but let me try it. I guess it is basically accepted that the validity does not lie in the instrument itself, but in its uses. I wonder if it is possible to make a distinction between psychological uses and program evaluation uses.

I think in terms of Dorans and Holland's work on population invariance, very interesting and important work in terms of coming up with tests that say, you give an individual a choice, do you take test A or test B, presumably they want to take the easiest test or what have you, and if there is no difference, then that is good for them.

But I wonder in a program evaluation context whether you need the same level of precision. For example, your point about the SAT and ACT having intermediate importance, but anyone who looks at those two tests would agree that they are indeed different tests, but they correlated over .9. In a lot of variables that we lesser social scientists are able to construct and measure, what have you, we would be thrilled with a reliability of .9.

So it may be that for our purposes, we don't need such stringent standards.

DR. LINN: I would agree that the standards that

you need if you are going to be making an admissions decision about an individual applicant for college, that the precision you need there is greater than you need for some other kinds of comparisons.

I guess that is why I would argue that looking at something like what Don has done or Henry Brown has done, and thinking of it as giving you approximations could have utility, even though it would not stand up to a close look, is this really equivalent in all the usual senses.

PARTICIPANT: Bob, could you discuss inferences between comparison over time of same grade level versus change from third to fourth to fifth, in terms of linkages?

DR. LINN: Well, linkages that you would get for NAEP, it is only going to happen at certain grade levels, so you're not going to be able to look at the vertical scale without this big gap between fourth and eighth grade, obviously. But if you thought of equating two tests that did have vertical scales, I haven't actually thought about this, so this may not be wise to say, but it seems to me that you might very well find that the implications about the vertical scale would tell you that the vertical scales are actually quite different. So if you equated the two different scales at fourth grade and fifth grade and sixth grade, you might have quite different equating functions, or noticeably different equating functions, because the

contents of the tests are not going to map in exactly the same way.

Steve Dunbar was talking about that indirectly yesterday when he was talking about how as you go up in the grades, what it is you are trying to measure is shifts, and the different test publishers don't shift in exactly the same way, I don't think.

PARTICIPANT: Of course, the unanswered question here is roughly speaking how good are the various state tests. You are giving us an excellent analysis of the extent to which they relate to each other. There is the remaining question of how good they are both technically in terms of content validity, for example, and in terms of the plausibility of the cut scores for proficiency, et cetera, all of which could be independently estimated.

If we are going to think about using these results for program evaluation, we have got to get some sort of a grip on how good or bad the various state tests are. It is a matter of some interest to the legislatures, I think.

So is there any work corresponding to Dorans and Holland's studies on the validity issues?

DR. LINN: Not that I know of. What states are putting together for the peer review process that we talked about a little bit yesterday, the peer reviewers are

looking at the issues of the validity of the tests. Bill [Schafer] could say more about that than I could, because I'm not a peer reviewer. But certainly the technical quality of the tests is in the minds of the peer reviewers and the state people who are trying to put together the evidence to pass muster to get their tests approved.

It would be nice to have all that information compiled in some way and made not just for the decisions that have to be made by the Department of Education, but for the evaluators and researchers to be able to look at the basis of the conclusions that peer reviewers came to.

Bill, do you want to comment on that?

DR. SCHAFER: Yes, thanks for the lead-in, Bob. Actually, I think there is a wealth of information that is available on these topics, if not in state websites, at least available and probably available by request from states.

I think you will find that virtually every state, for example, will have alignment studies that are done by external agencies that conform to standard procedures, and those agencies produce reports. Peer reviewers review those reports regularly. Technical adequacy is a fundamentally important area in peer reviews. Again, there are technical manuals that accompany the tests, so the technical quality is something that can be documented.

DR. LINN: Thank you.

DR. DUNBAR: I am going to suggest that we hold any remaining questions for Bob until the general discussion, and move on with David [Thissen]'s remarks.

DR. THISSEN: I'd like to thank my friends at NRC for inviting me up so I can see some snow before Christmas. I'd like to thank Bob for a great paper to allegedly discuss, but actually to use as a jumping off point, because it taught me things, as Bob Linn's papers always do, and provides a lead-in to a few comments that I will try to make as quickly as possible to keep us on schedule.

That will be made somewhat easier, by the fact that I got this ready early this week, and then yesterday various someones said most of the things that I have in mind to say. So I will try and do this to the extent possible with the cross referencing to yesterday. There is a written version, and since it was written before yesterday, it doesn't actually have the cross referencing to the things that other folks were saying.

The database which is the victim, I mean subject, of the workshop, invites comparison across states. Anything in the United States, it seems, that has numbers for more than one state invites comparisons across states. A significant obstacle to that, which was mentioned extensively yesterday, is that all states use different

state assessment systems, and they have different score scales.

Bob's presentation reminds us that the idea of trying to make comparable scores from tests that differ like that has been happening. The idea has been circulating for at least 40 years. Test theorists have been fairly negative about that idea for at least 40 years, and practice has tried to ignore that negative for many of those years.

If one is concerned about the lack of comparability of scores across the states, one answer is to do only within-state analyses, which Bob didn't go on so much about in his talk, but was the first thing that his paper lists as an alternative.

Setting that aside for the moment, Bob talked in his paper about two strategies for combining results across states. One was using standard scores or effect sizes, which wasn't so salient in the slides that Bob just showed, but it is in the written version of the paper. The other is about linkage.

This leads me to doing the discussant thing, which is to cherry pick, to pick and choose topics to go on about, possibly to do some pet peeves. So I would like to talk a little bit about using standard scores or effect sizes in these concatenations across states, and then talk

a little bit about linking to NAEP. After that, I hope I will leave myself time to say some more general things about what one might be able to do with these data or data like them.

Before I read Bob's paper, because it was before it was written, I looked at some of the materials that were circulated to the workshop planning committee, papers that had used the database for various purposes. I noticed that in the process of making standard scores or effect sizes, that often it seemed that because the database includes school level data, that in order to get standardized scores or effect sizes, the way that was being done was to divide by the standard deviation of the school means.

It occurred to me that that might lead to some unintended interpretations under some circumstances. I turned out not to be entirely alone in this idea -- Judith Singer went on for awhile about this yesterday afternoon for those of you who were there -- but I am going to tell briefly my story about it.

It seems that the standard deviation of school means includes at least four things, and I say at least, because I can always miss things, but when I thought about it, I could think of four pieces that it includes. It obviously includes variation among students, which Bob has pointed out privately on occasion cannot be legislated

away.

Aside from that, another couple of components are size of the schools and the sociology of assignment of students to schools, as well as school effects. That is what people are concerned with in the analysis of this database.

But letting that trail away a little bit, my old friend Howard Wainer visited the Carolinas a couple of weeks ago and was giving talks, one of which had this extremely large equation. He was talking about school size as a variable. He had the slide that only had the very large equation on it. I copied it, but decided to add some words.

This is from the first course in statistics, and it reminds us that the standard deviation of school means becomes smaller as schools become larger. Why do we care? What I think should be the same in terms of an effect size is the same difference in the kind of units we usually use to score tests, which are usually referred to the standard deviation of students, of some reference population.

Now, the same difference on that scale becomes larger or smaller if you divide it by a larger or smaller standard deviation. The school level standard deviations become larger or smaller simply by making schools smaller or larger. This varies between states. Some states have

larger schools or more larger schools than others. So if you do divisions by school level standard deviations, you could get differences between states that in some sense, with some other standard deviation, wouldn't really be there.

The third component is particularly salient to me. I don't know how many of you live in communities like I do, Chapel Hill, North Carolina. The way students are assigned to schools is fascinating. At least in my home town, the newspaper reports it every year, and the school board goes to a great deal of trouble gerrymandering school districts, busing people, trying to arrange the assignment of students to schools so that the students in a school are as heterogeneous as possible in a town full of college professors.

The effect of that is to make the school means as similar as possible, so real estates don't vary particularly much within the town. But this kind of effect on the school mean makes it have less variation than one might think.

I thought of another -- I think my long sentence from the ETS conference may have been incomprehensible, so I got a shorter rhetorical question for this one, which is, what if the standard deviation of school means is nearly zero. That doesn't happen for a variety of reasons, but it

is an interesting gedanken experiment. That would mean you would divide by zero and your computer program would crash if you did this.

How could that happen? You get a state that has very large schools, you assign the students to schools to maximize within-school variance and minimize between-school variance, and then have either no school effects, or school effects that counterbalance the effects of background variables, and you could wind up with very little variation between school means.

This has been a long pet peeve, which expands a little bit on things that Judith Singer was saying yesterday. A student-level standard deviation which you can get for most of these state tests would be a better denominator, I think, for standardization.

But now we get to my next comment, however, which is going to link into the second bit. Even if you standardize the tests in some way that is referred to the student scores, you may not get comparable results. This is what Bob talked about quite a bit more extensively, using linkage to NAEP to try to make these non-comparable scores relatively comparable.

My story about this -- these are the covers from the committee reports, which are colloquially entitled No Sub-1 and No Sub-2, that Bob talked about. I got involved

in this because in the 1990s, we did one of the linkages of a state test in North Carolina. I thought it was cool at the time. We used some interesting statistical procedures. My memory is that the first time I spoke about that in public was at a conference at NISS, the National Institute for Statistical Sciences. My vague memory was that Don Rubin was the discussant. After I said I thought this was cool, he raised some very serious questions about whether it would be invariant across subgroups that we hadn't checked. That was mildly embarrassing.

But it turns out that after some thought, these reports raised some questions about NAEP can actually be used to line up disparate assessments. You can read the books, but I'll do a couple of short versions. This will get to the issue of, does this mean something for program evaluation, I think.

There is some level at which grossly one ought to be able to line up one math test with another math test, but where grossness ends off and you need to be reasonably precise can easily get into the range of material you might work for even for aggregate program evaluation.

Here is something that could happen. States A and B perform about equally on NAEP. State A does better on its test. If you hypothetically gave that test in both states, state B does better on its test. The reasons given

for this were mentioned yesterday. Stephanie Stullich's presentation talked about curricular and blueprint differences. Laura Hamilton talked about different stakes. This is something that could happen.

But to get away from something that could happen to something that did happen, what you might want to do to check equating state tests to NAEP or linking state tests to NAEP would be, have some states that actually gave the same test, where students took NAEP, and see how that came out.

Now, because states use different tests, that almost never happens, and it didn't really happen here, either. There are four states over here, anonymously called states one, two, three and four, which in 1990 used as their statewide assessments four admittedly different tests, all of which were published by CTB in California. If my memory serves me, it was three different versions of CTBS, and one was still using CAT-5.

However, those tests had been in-house linked to each other, to the CAT-5 NCE scale. Courtesy of Bob's talk, we all know what NCE scales are now. I had never known myself. In any event, the population state means for the four states are quite close. State one is highest, the other three are very closely clustered.

Now, in 1990, NAEP TSA was administered in those

states. State one was highest, but state two, which is tied for second over here on its statewide test, was lowest by a chunk on NAEP. State three comes across the middle, and state four dips down to be over there. So you have got states over there and over here, and they are not really very close. They are not in the same order, they are not in the same spacing. The spacing is sort of meaningful.

This is the kind of result that led to the conclusions in the NRC reports entitled No Sub-1 and No Sub-2. It leads to the conclusion that there is no guarantee that linkage to NAEP or anything else can make comparable scores obtained with different state assessments.

That was the pet peeve portion of the discussion. I would like to get on to saying some positive things, because in discussing the symposium in advance, we were told we could make suggestions about what might be better, what could be added to the data set to improve things. So I put together a little bit of an overview of what I think we would like to have in order to do the kind of program evaluation that is discussed here. We would like to have student level achievement scores on comparable measures reflecting progress toward the same curricular goals.

Actually, David Goodwin's introduction to the whole workshop yesterday basically said this. The slide

may be plagiarized, but I did it in advance.

What are available are non-comparable measures of different curricula. The first thing I would put in this, which has been coming up and coming up and coming up, was Judith Singer's most important point at the end of her discussion yesterday, which is that student level results in a database would help.

Many states are at this point in developing student level databases. Last night we had a fairly extensive discussion of how this could be done. Even though I was taking for the sake of argument the negative, I was pointing out it would cost a great deal of money and would take creative solutions to the privacy problem, we were pretty much in agreement that this could be done, and would take away a whole lot of the difficulties with using this database for many of these purposes that we have been talking about.

The lack of comparability of achievement tests across states is another matter. At the very least, that is difficult to solve, most likely it is impossible to solve. That may not be so bad. This doesn't stop anyone from doing analyses within states.

Cross-state comparisons may add relatively little information. If one wants to do cross-state comparisons, we talked about using meta-analysis to combine the results.

Using meta-analysis to combine results when you have all of the data seems a little bit odd. Meta-analysis is a set of techniques built to work around the fact that the meta-analyst only has summaries. If you have all the data, you can simply analyze all the data.

However, there is a time honored tradition in meta-analysis called vote counting, which doesn't depend on metric comparability of the results, and might be a way to combine results across states that wouldn't require standardizations that lead to discomfort or links to NAEP.

This is going to end. This is the last slide. I was both timid about it and a bit proud of it when I made it, because I thought I was going to say something very strange. The little recipe on the slide suggests an approach to the analysis of these kinds of data which was not actually the kind of approach I was seeing in the papers that had previously used it that we were given in advance. I thought I was going to be off the wall deviant in saying this until Judith Singer did this whole bit much better than I am going to do it, for half of her talk at the end of yesterday. So now I am in this awkward position of saying, "me, too."

But the idea is that instead of approaching this database with the kind of data analysis that says, "what we have in the database are school means for these states, now

what can we do with them to answer the question?", we will have to adjust for this and make that correction, and combine in this way as a sequence of data analytic steps. We could do something rather different, which is to ignore the data for awhile, to develop what would probably be in these contexts of program evaluation for educational interventions, to develop a student-, school-, maybe state multi-level model which would be the model that included parameters whose estimates would answer the research questions you have. That is, make the model to give you the answer you want, never mind the data. That is step one.

An example of such a model would be the kind of models that Yeow Meng Thum talked about yesterday. I will leave those as examples, because other folks make multi-level models too, and I am a test theorist and I don't want to stand up here and get myself in trouble endorsing one or the other. But that would be the kind of model.

Now, you can't estimate the parameters of one of those models from the data in this database, from the aggregate data, but you could figure the implications of the model for the aggregate data, and then you could figure out what you could estimate and what you couldn't, at which point you would have two choices. One choice is, you could say that with these aggregate data we can't really figure

out enough from these data to make it worth it, and you stop. That would be one option for an end. The other is that you could figure out what assumptions you were making very explicitly about what Judith Singer referred to yesterday as the missing levels. The student level would be missing. The inter-class correlation among students would be missing. You could figure out if you could get sensible answers without knowing those, and go ahead. But in the process you would have made very explicit what assumptions you were making in order to answer the questions with the aggregate data.

As I say, this is a "me-too" thing. If you want to see a better description of it, probably go for Judith Singer's discussion.

That was the last slide, and I think I better stop. Thank you very much.

DR. DUNBAR: Any brief questions for David? Why don't we move right on? Don.

DR. RUBIN: I was really glad I got invited to do this. I have been more or less disconnected from the education community for awhile, and it is nice to see a lot of faces that I haven't seen for actually decades. Michael [Scriven], the last time I saw you must have been three decades ago, is that right? It is nice to see that we haven't changed, either. That is always good news.

When I was first contacted about this conference, I had the feeling that it was a focus on causal things. I have been doing a lot of causal inference in the last couple of decades. Sometimes I work with medical people, sometimes with economists, not as much with people in education. And of course, I was sort of turned off with the idea of using these very observational databases to have much chance to doing anything really in terms of evaluating programs. The challenges to doing it are so great for data that aren't even remotely collected for that specific purpose, to do evaluation of interventions. It doesn't mean it can't be done, but Dave [Thissen] commented at the very end that what you can do is, you can say what you can do, and then add statements about what the assumptions are that you have to make to add to the data to reach some conclusions that you might want to reach is an important idea.

In other words, to draw an inference you always need some assumptions, especially when you are drawing causal inferences. You always need some assumptions, and you are fortunate if you have data. The better the data are, if it is a joint randomized experiment, the fewer assumptions you have to make. You still have to make assumptions. But the worse the data are, the more you have to supplement the data with assumptions.

Sometimes the assumptions are fairly plausible. Certainly we accept all sorts of conclusions, causal conclusions, for which there is no randomized experiment to ever justify it. We are happy to do that, and they make good sense.

The one example that comes to mind, we all believe -- we should believe, at least -- that cigarette smoking is a cause of lung cancer. If you smoke cigarettes, three packs a day, you are more likely to, causally more likely to get lung cancer than if you don't. Yet there has never been any randomized experiment that supports that. There has been no randomized experiment with human beings, but in fact, there has been no randomized experiment with animals. They have tried. They have tried to do lots of randomized experiments with rats, with dogs, where they are forced to smoke by having cigarettes in a smoking environment for their whole lives. They don't get lung cancer. Sometimes they get other kinds of cancers like skin cancers, but they don't get lung cancer. Yet we all believe and should believe that smoking causes lung cancer.

Why do we do that? Why do we believe that? One of the main reasons is an analysis like the one Dave just mentioned, which is called sensitivity analysis. At the time that the smoking controversy was rising in the mid to

late '50s, the tobacco industry had a consultant named R. H. Fisher, who is also a famous geneticist, who was claiming that it was probably some common unnamed genetic factor. If we could measure and adjust for this factor, the relationship between smoking and lung cancer would disappear.

What Kornfeld did, who is statistician at the National Cancer Institute, he did what is called a sensitivity analysis which was very much like what Dave was doing. He said, let's suppose there is an unmeasured factor, some other genetic thing. How strong would that factor have to be to make this observed association between smoking and lung cancer disappear? It turned out that the relationship would have to be much stronger than any other genetic factor that has ever been around. It would have to be so strong that Fisher and others classified it as an incredibly strong genetic factor. So at that point Fisher was silent, and people accepted that it was probably true.

Logically, what is true? Lung cancers take 30 years to develop, typically 20 years, 30 years to develop, and rats and dogs don't live that long. So it is a slow-growing cancer.

But anyway, it is in support of Dave's very important point that if you are going to use data like these that are not collected well for the purpose of doing

causal inference, don't just use them, but use them intelligently, which means to me doing exactly what he was talking about, which is to say, here are some conclusions based on data. In order to make them causal, we have to add all these assumptions and carefully think what those assumptions are, and try to look for evidence supporting those assumptions somewhere else.

Now, the other topic that was more the focus of Bob's presentation was test equating. It brought back memories of a conference at ETS that Paul Holland and I organized probably close to a quarter century ago, which ended up in a book that Paul and I edited together. I think it was entitled Test Equating. Academic Press published it in 1984 or something like that.

But I have thought on and off about some of those issues since then, but not that much. At the time of that book, I was at ETS, and Paul and I were actively involved in trying to do equating for ETS. At the time there was a change in New York state law, some truth in testing law, that you had to release the items that were used to produce the grades on the test. Before that they kept this item pool, and they kept it secret. They never told the kid which question he got right and which question he got wrong. But after this, at the end of a test you had to release all the items. So ETS had to decide to either get

out of the business of testing in New York -- but that made no sense because other states were going to follow -- or to change their way of equating the test. They couldn't take a pool of items and randomly sample within item pools.

One of the things that Paul and I proposed and did was something called section pre-equating. I think it is still used at ETS. It was based on missing data technology that was just in the forefront of statistics at the time, a lot of it based on something called the EL algorithm, which has become very, very popular in recent years. It was in a paper that was written in 1977.

So we used that. One of the key ideas there is, when you think about equating these tests, you think about what would I really like to have if I could get all the data? What would all the data be? David, in your written version of your talk, I think you may have mentioned this, maybe not, but all the data here would mean if each kid in every state took every state's test, and they took that other state test first, so there are no practice effects of taking the other test. This is obviously impossible. You can't take every test first. But it is a way of thinking about what the ideal data set would be, and then from that you can do everything. Certainly you can do everything that anyone is talking about doing, because that is a super-big data set.

How far off is our existing data set from that?

The question here is -- the discussion has all been on, the existing data are there anyway, so what can we do with them that is worthwhile. That is a passive attitude in some sense, the data are there, what can we do with them, and that is okay.

And of course, in that view what we have got is a big block diagonal 50 by 50 matrix. Across the columns are the 50 different state tests, and little sub-columns in that are the different items on the tests. The rows are 50 blocks of rows, or each row is a state, and within each sub-row is a kid who is taking that test. So you get this matrix of existing data. They aren't like the ideal data because all they are are the diagonals. You've got individuals in the state taking that state test, running down the diagonal.

There is some other background information that may be common across all the states like the NAEP stuff. So that sits out there as a giant covariate, that is fully observed on everybody.

The question is, what can you do with this big matrix where you can only get block diagonal. Maybe there is a more active proposal, something that may be far out. Properly deserved so, a reaction from most people in the audience and maybe everybody, is that you are nuts to

propose this, it couldn't possibly work. But let's think about what we could do to enrich the data set so that maybe we could get almost all the answers that we could get in this ideal data set, if everybody took every test first.

That is why I am cycling back to this idea of section pre-equating. Let's suppose that you took a small section of each state's test and took a small sample of each group of kids within each state, and gave them the test, give them different sections of the test. So you had all these different test forms, hundreds and hundreds of them, but every section of each state's test appears once in one subgroup of kids with every section of every other state's test. So you have these overlapping pieces.

Why do they have to be overlapping? Right now you just have the margins. You have Ohio's test in Ohio. You get to estimate the mean and variance in that state. But maybe with these overlapping tests you get to estimate these correlations as well.

By doing the proper sampling of who is going to take these sections of tests, you make sure that they are random within the state. That would give you an opportunity, of this giant matrix where we just have existing data down the diagonal, you would now be supplementing with little pieces off diagonal.

What can you do with that? It turns out that the

same missing data technology that is being used to section pre-equate tests can be used in principle to get all the means and all the variance and all the co-variance and all the correlations among all the different tests, and all the different subtests. It is a gigantic computational problem.

When these algorithms were proposed in 1977, everything was a giant computational problem. I remember when I was an undergraduate at Princeton in the '60s, the mainframe computer that filled up a room almost this size, the memory was 28,000 words. Wow, was that huge. Probably 90 percent of you have these little pocket things that have 50 times that much memory. The only way Princeton could afford it was it was bought by National Energy for doing things for the fusion projects. It cost millions of millions of dollars then, the equivalent of hundreds of millions of dollars now. Computing has gotten really cheap, so the computational burden to this isn't that much.

What is the real burden? The real burden would be making up the test forms, doing the distribution in a way that it would be approximately random, getting all the states to agree to give these different subtests.

The idea is that you take the regular test and you add something at the end, a small section at the end, it may make the test 15 minutes longer. So each kid

doesn't have to do that much more than they did before, and each kid is doing something different. This idea of ordering is the way it was done 20 years ago at ETS. You take this section, you don't really put it at the end, but you put it in the middle of the test, so it is an equivalent section with some other section, and sometimes it is practiced, sometimes it is not. So you counterbalance that out of the procedure.

The assumptions under which this stuff works, you are relying on normal distributions to some extent, but tests pretty much are. Even subtests pretty much are. You don't really need normality; you need something more general called ellipsoidally symmetric distributions, where you have linear regressions and stuff. So it could work pretty well. It works pretty well for section pre-equating, I think.

You have to define what these subsections are and how they might be common across different states' tests, but it gives you a tremendous lot more information, because in fact, what it does give you is, it gives you an estimated ideal data set that I talked about before. This ideal data set is if every kid took every test from every state unpracticed. That is what you get to estimate by this. That is a pretty nice hunk of data. And because there are so many kids in so many states, the burden per

kid isn't much. The burden per state for getting the information, yes. But if you really want the ideal data set to have, this is the way to go after it.

So that is my suggestion. I am prepared now to get attacked. Maybe I should say, because I was here in spirit yesterday, it does also suffer from Judith Singer's comment that you need individual level data. Here, when you are doing the equating part of this, when you are doing the subsections, you need individual kid data to get the correlations.

You don't need individual kid data on a lot of the data. On a lot of the data you can just live with the means and variances, but for some subsections of the kids you need individual kid data, and maybe that is okay. Maybe that is all you need. Maybe you can get it. It is a question of, is the benefit from having this data set, the kind of analyses that you could do from having it, worth the extra effort to get the extra data. I'm not an expert in that at all.

Thanks.

DR. DUNBAR: Despite our scheduling, our speakers have more than cooperated in giving us still plenty of time for questions and discussion. So we will open up the floor for that.

Is it more than a modest proposal to suggest

augmenting every state's test with 39 others, plus?

DR. CHESTER: Mitch Chester from the Ohio Department of Ed. It is a great proposal. Probably the hurdles to get there are huge, and the incentives, and so forth. But it does make me wonder whether there are naturally occurring situations, such as kids' mobility between states, where if you could link the records, you would have a lot of issues around whether or not you have got representative samples and so forth. But maybe there is a potential where there are some naturally occurring sets of students who are taking multiple tests, multiple states' exams.

PARTICIPANT: If you actually wanted to think about doing this, it would seem like NAEP would provide a framework where you are already sampling students in every state, and you already have blocks that you are counterbalancing. You could probably do the math and see how many different blocks you would have to have to represent how much content in how much state.

The requirement that all states participate in NAEP has led to samples that are more huge than NAEP knows what to do with, so there might be samples there to use up for this purpose.

I did also, related to this point, want to point out that Don McLaughlin had done an earlier study where in

four states they got individual results on NAEP linked with individual results on the state assessments. It shows variation in correlation, but that is the kind of data that allows you to begin to answer some questions about how much of a difference did the violation of various assumptions make in the way in which students or schools would be ordered by different state assessments.

PARTICIPANT: This relates to the first two presentations. Bob Linn talked about NAEP linkages. In your paper you were using proficiency scores, basically. Dave [Thissen] talked about NAEP linkages. You were talking first about school means and then about going to student means.

But to raise the issue that was raised yesterday, what kind of advantage do you get in making NAEP linkages by moving from proficiency scores to parametric data, mean scores and so forth, which we know is not feasible quite yet, but may be at some point.

DR. THISSEN: Well, I was talking about the scores and the means because I defaulted that. There is always more information in the scale scores than there are in the categorical data. In all these tests the categorical data are just obtained by dividing the scale score range into chunks. So I think that where one can, if one wanted to do the kind of development of relations among

tests that Don suggested, one would tend to do it with the scores.

Bob and I will put words in his mouth and he can take them back out. He was talking about these analyses that Don McLaughlin did, the very early ones, answering this question of whether NAEP can be used to shed light on what is in some sense the very obvious fact that cut scores for what is called proficiency are extremely different from state to state. That is the reason that 80 percent of the students in one state are called proficient and 20 percent of the students in another state are called proficient.

At the very gross level of the difference between 80 and 20 percent, I see no problem with using NAEP to draw the conclusion that the NAEP cut points are different. But I also showed you where NAEP means are differently ordered than state means, which one would hope are on equated tests. So at some point you go from gross differences that you can infer from NAEP are not quite stacked up right, that you can infer from NAEP are due to the tests being different, to somewhat finer differences which probably are due to things like the fact that the tests measure different things, and motivational circumstances are different and so on.

Don [Rubin]'s proposal of this extraordinary data collection -- and it would be extraordinary, I certainly

never thought of anything that extraordinary. But I think that a difference -- this is not by any means a showstopper, but the merger of the kind of models that I was talking about at the end of my presentation that Judith Singer introduced yesterday, and Don [Rubin]'s presentation is that part of those models would, I think, have to deal with the fact that the different state tests and the NAEP don't really measure the same aspects of proficiency. There is this broader and narrower kind of aspect of achievement tests that Steve talked about yesterday. There are differences in curricular emphases.

So the model would have to be somewhat different than section pre-equating, which was built for forms of tests that are supposed to be parallel at the end. That is not a showstopper, that is just more modeling. But it would have to be explicit in there that you are not actually making a table that -- if you produced something that looked like comparable scores from all the states, you would have to be in your analysis making up the combination of achievement that this superordinate measure measured. It wouldn't presumably be exactly what NAEP measures or exactly what any of the states measure, but rather literally a construct, meaning something the data analysts constructed. You could do it, but it would be an adventure.

PARTICIPANT: Your last point, that is something I have been thinking about quite a bit as well. Yes, we have the ideas of pure construct in our heads most of the time, but how do we normally deal with it? We feel around using various tests to get at various aspects of the construct. We never pretend that a test set out to measure math is exactly that.

So to some degree, again "me too," I will follow, and a more pragmatic look at the whole situation. While we can work on statistics that show deviations, population invariance, we should be continuing those studies across time. It is not one set of stuff.

What do you normally do when you have differences over time? You average. The linking questions, we could perform some averaging over them and basically say that this is the best we can do, on a construct that is approximated as best we can right now.

So I am operating on a so-called satisfying principle in a complex decision making environment, involving instruments we have to construct. So I couldn't think of a design that Don [Rubin] came up with, but it really sounded interesting. It leaves aside the possibility for averaging over all the different scales that come up with something on which everything is weighted in some way. All we can say is, it is math of some kind.

PARTICIPANT: I want to add that I completely agree with Dave [Thissen]'s point. This reconstruction of this hypothetical giant matrix doesn't depend upon the section pre-equating assumptions. It is what you do with it that depends on the section's pre-equating assumptions. All it is doing is saying, "let's suppose that underlying this data is a giant multivariate normal distribution, and I am going to fill it all in. I am going to get all the means and variances and correlations, and assume they are related, that's all."

Now, if you then do something with it in some section pre-equating way, that is right, you are making some sort of assumption like that. I completely also agree that the kinds of things you would like to do with it are to say, this state had a good idea here, this state had a good idea there, this state had a good idea here, and pick and choose the subsections of the different states and make a hypothetical giant test that nobody took, but everybody has parameters for it that are state level parameters, and you then get an answer for that test, an estimate for that test, that nobody took. And you can change the test, depending upon what you want to ask.

So it has virtues like that. It is an idea that has a lot of richness to it, I think.

PARTICIPANT: I understand how engaging in this

enterprise would help us to understand better how the tests in different states are comparable or non-comparable to each other. But from the standpoint of the program evaluation of federal programs, presumably the programs are implemented in such a way that we get a measure of the effect of the program in each state.

So from the standpoint of understanding program effects, the value of the state test and differences in the state tests is understanding heterogeneity in those effects across states. So what I am not clear about is why we need to know about the comparability of tests across states to understand those effect sizes for the program effects. Can we use current knowledge about differences across states to look at heterogeneity in the effect sizes as a moderator variable?

DR. RUBIN: I guess there would be two answers to that. One is that this proposal is just to address the construction of this ideal data set, and not say exactly how you were to use it.

The issue of effect; effect is a tricky word. Do you mean causal effect, like the effect of an intervention? Or do you mean an effect like, "I did a random effects model and each state has an effect, or each school program has an effect?" Three different things.

PARTICIPANT: Let's talk about the causal effect.

DR. RUBIN: Going back to the first comment I made, you really are dead doing that unless you bring in a huge number of assumptions, which is okay, but you better explicate what assumptions you are making in order to get there.

For that, I agree. If you are only looking within a state and say, we just implemented program A, and now we would like to know whether program A for the state works in the state. You don't have other state data necessarily, but let's suppose you did, so that you wouldn't only have before and after A in that state, but then you had an adjacent state, Virginia and Maryland or something, where Virginia didn't do it. So now you get more information to try to calibrate what happened in Virginia because Maryland didn't change.

There would be differences which I don't like, but the idea is that even if you are only interested in one state, what is happening in a neighboring state where there was no change, can be useful.

This was used, for example, in a study by [Alan] Krueger and David Card, which evaluated the minimum wage. So they went into Pennsylvania and New Jersey. New Jersey changed its minimum wage and Pennsylvania did not in one year. They looked at Burger Kings and McDonald's as restaurants to see what the unemployment rates were like in

the two states before and after the change, using the state that did not have the change in minimum wage.

PARTICIPANT: So in those situations where federal programs are not implemented nationally, but in certain states and not other states, then comparability could help.

DR. RUBIN: Yes, because wages meant the same thing. If you paid people in lira in one state and dollars in the other, you first have to go around and say how do I convert lira; it makes it harder. So if you had this giant data set, it gives you a leg up on addressing those questions, even if they are focused only on one state.

PARTICIPANT: (Comments not picked up by microphone.)

DR. RUBIN: For that, I think it would actually buy you even more, because you get to define the outcome test that you wanted, not corresponding to any state's test, but to some composite test that no one ever gave anybody, which would be considered to be fair across all the states perhaps. Then you can use units in states and kids within states, and you would have a lot more structure to try to build on.

PARTICIPANT: In order to make these comparable inferences, for example on effect sizes, would there have to be a uniform methodology for all the states to make a

determination of effect sizes? Would that then have to be established by the Department of Education and CES and other commonalities, so that you have got different measures, maybe different achievement levels, different ways of setting achievement levels? But then the statistical analysis, would that have to be uniform in order to make these inferences across states and across tests?

DR. RUBIN: I don't see why. The thing that has to be carefully designed is the sub-samples of kids who were taking the state and the sub-samples of sections of tests. But they would be given in the same way that the regular states' tests are, because that is part of the action right now. So I don't think you want to actually change the individual states' tests, because that would be too hard.

PARTICIPANT: (Comments not picked up by microphone.)

PARTICIPANT: But we were trying to build the database from raw data, and then doing all the inference at some uniform level. So you're right. I think I probably misunderstood what you said.

Once you create this database from the states' raw data that they give back to you, then the analysis to build the database and the analysis of it would have to be

uniform.

DR. DUNBAR: One more question.

PARTICIPANT: Can you explain how this is different from No Sub-2?

DR. RUBIN: If I knew what No Sub-2 was, maybe I would be able to.

PARTICIPANT: Maybe David [Thissen] can.

DR. THISSEN: See, No Sub-1 and No Sub-2 -- well, No Sub-2 was talking to some extent about using data designed not quite as elaborate as Don [Rubin]'s. Don's is really big. But what No Sub-2 said is, you couldn't take that [comments not picked up by microphone] -- I guess what wasn't thought about was this idea that we evolved to up here in this discussion of making up a score scale that is referred to a construct that doesn't actually match any of the above. That is, take math; it is not math as defined by NAEP's blueprints and items. It is not math as defined by ITBS. It would have to be either something else or possibly even a multivariate something else.

I don't think the committees of No-1 and No-2 considered that. What they said is, you wouldn't get comparable scores if you just made one to one concordance tables between scores on NAEP and scores on ITBS, because that wouldn't give you the same shufflings of many answers.

We may have actually come up with an idea of that

out of all those committee meetings; in my memory it was something that never actually came up.

DR. RUBIN: Does that mean I get my expenses?

DR. DUNBAR: One more question from Laurie Wise, and then we're going to have to stop for lunch.

DR. WISE: (Comments not picked up by microphone.)

DR. DUNBAR: That's right. Well, on that note, I think we will close this session. Join me again in thanking all of our speakers.

We will break for approximately one hour. We reconvene at 1:20 sharp.

(Whereupon, the meeting was recessed for lunch, to reconvene at 1:20 p.m.)

A F T E R N O O N S E S S I O N

**Agenda Item: Session 4: New Opportunities**

DR. KINGSBURY: Good afternoon. I'm Gage. I work with the Northwestern Evaluation Association.

I was asked to talk a little bit about some of the ideas that might be used to bring the SSASD database a little bit closer to the ideal, from the point of view of program evaluation. So for today, I am going to take off my psychometrician's hat and talk a little bit with my evaluator's hat on. Psychometricians' hats are notoriously hard to put down, so I will probably pick it up as we talk.

I'll talk a little bit about the strengths of the database as it currently exists, and then I'll talk a little bit about some of the weaknesses it has for program evaluation, and then finally I'll talk about some ideas that might be used to move it a little more in the direction of the ideal if not to the ideal. Then I'll finish with an example of application of one of those ideas.

The strengths of the database are pretty clear. There is centrality, a broad variety of data brought to a common location so researchers can get at it. That puts us ahead of where we were ten years ago pretty dramatically.

There is substantial breadth in the database. Even though not all students are included in the testing in

individual states, enough of them are included to make it interesting for use in a program evaluation standpoint, even when you are trying to evaluate a program that is a wide-ranging federal program.

The database currently has state level consistency. I'll talk a little bit more about that because it goes to the question that was raised this morning, not suggesting that there is any consistency from one state to the next to the next. But for a change, we have gotten to the point where there is consistency within a state and a fair amount of data within a state that can be used to look at how students perform from one grade to the next to the next.

It is not as crystal clear as that. There are still a fair number of states that use multiple test vendors to provide their tests at different grades, and as a result, in those states the consistency isn't clean, but substantially better than it used to be.

There is a motivational consistency that has been lacking in the past. Let me tell you a little bit about that. The problem that we have had in doing evaluations since the dawn of time is that when you are trying to do an evaluation study, you have to coerce people into doing things. One of the things that you almost always have to coerce people into doing is taking some sort of pre-post

measure, or some continuous measure so you can identify change.

It is not always the case that the individuals taking those assessments are the most motivated people in the world. They are not always the least motivated, but probably closer to that end of the spectrum. So as a result of No Child Left Behind, we have at least relatively consistent motivation across states, and certainly decent consistency of motivation within states. I'll get back to this whole issue of motivation in a little bit.

The last thing that we haven't had before is longitudinal information. It hasn't always been easy to capture longitudinal information, particularly to deal with the kind of time series analysis that starts before a program goes into place and follows implementation and then asks questions after it has been implemented. The longitudinal information that we get from this database and that is caused by No Child Left Behind enables us to ask questions using data from today for a program that won't be implemented until tomorrow. That puts us in a very strong position with respect to time series and things like that.

Enough about strengths. I don't like to talk too much about strengths; it wears me out.

There are some substantial weaknesses in the database, and they deserve to be thrown out so we can

figure out how to fix them. The first one isn't actually a weakness of the database, and that has been brought out fairly quickly and fairly completely by Michael Scriven's talk yesterday.

What we can do with this database is look at longitudinal data and not create random control design experiments, at least not using backtracking data. I kind of dislike the term quasi-experimental, just because I think it suggests something that most longitudinal designs aren't. So I suggest eliminating the words quasi-experimental from your vocabulary if you can, and consider a longitudinal design which, rather than being quasi-experimental, is a process to ferret out the truth, just as an experimental design is.

So if we think about a longitudinal design, and we think about it creating evidence beyond a reasonable doubt, I think was the term someone used earlier, that is probably a good place to stand. But somewhere in this process, the parts of the Department of Education might want to talk to each other so that we can get to the point where longitudinal data analysis isn't considered a second-class citizen in terms of identifying causal factors that enable student growth.

One of the things that you always like to see as an evaluator is an indicator of how well and to what extent

a program is being implemented in a particular site. So if you are doing a multi-site analysis, you know that site one implemented reasonably well, site two implemented not at all but they had the materials, and site three implemented extremely well, because knowledge of the implementation is imperative to figuring out whether what you have is kind of a dosage or whether it is an issue that a program simply doesn't work. So that is extremely important to have, and unfortunately we don't currently have that in the database.

There is also a small problem with inconsistency before and after No Child Left Behind. Before No Child Left Behind, the data for state tests wasn't necessarily always captured in the most motivated of settings. It was very common for teachers to try to get their students to attend to the test by saying things like, "you will never get a result from this test, but do your best anyway."

That is not the best of all possible worlds. So we have an inconsistency before and after No Child Left Behind in terms of motivational factors. We also have inconsistency before and after the tests being used, in terms of whether or not there was a proficiency level established. A lot of things cause that difference, and those differences are going to play out in unknown ways when we try to do evaluations of programs.

There are different tests. You know that

already. There are different score types, and that is probably a useful thing to talk about, since everyone else has, so I feel compelled.

The different tests in different states have a variety of different score types that are available. One of the strengths of the database right now is that it has given us a common language for dealing with those different score types. That is a good thing. The bad thing is that not all score types are available for all tests, so some tests have vertical scales, some tests have scales within a grade, and some tests don't actually give scale scores, but drop things into buckets. So there is a substantial difference in the level of granularity and the type of analysis that can be done with different score types.

We consistently see people try to analyze percentile ranks as if they are score scales, and it is not necessarily the best thing to do, from either a statistical point of view, psychometric point of view or an evaluative point of view.

The last one I wanted to mention as an indicator of weakness in the database is that we don't have an indicator of differential accuracy. For the evaluation of a federal program, one of the things we are going to need at some point down the road is an indicator of how well a particular test does in measuring the performance of a

particular group of students.

So, for instance, if we are studying students at risk and we are using -- let me take an example from some work we saw this morning. We know that South Carolina has relatively high standards relative to the rest of the world. As a result, a test that is designed to identify whether students are proficient or not in South Carolina probably isn't going to be overly well designed to measure subtle differences between students who are struggling. So the psychometrics involved in the test may get in the way of our evaluation, depending on what state we are looking at, depending on what their proficiency levels actually are. It is something to keep in mind.

Unfortunately, right now in the database we don't have an indicator of the reliability, validity, information functions, or anything like that, that are associated with the individual tests. The data exists; it is just not neatly available for the evaluator to use.

Enough for being a downer. Let's talk about some suggestions for change, which is why I am here in the first place. One of the first things I would suggest is creating a score validity indicator. What I am thinking about here is, within any school in the database, there will be a certain number of students who respond to a test, and the test is either far too difficult for them, or the test is

far too easy for them. Those students tend to get scores that aren't the best indicator of the student's achievement, but we don't treat them any differently for the most part than we do students who are smack dab in the middle of a measurement range for a test.

So it might be useful on a school level to have an indicator that tells us what percentage of students in the school got a score that would be considered in the invalid range. I would call invalid pretty much anything that is at or near chance, pretty much anything that is at or near 100 percent accuracy, because once we get to those points on the scale, regardless of the psychometric model that we are using, what we have in the data is a substantial amount of noise compared to a relatively small signal. That is particularly true for the students who are performing at the lower end of the distribution. The amount of noise that gathers in a score for a student who is really struggling with a test is surprising.

So it would be useful to have a score validity indicator at least on the school level, which is our lowest level. Obviously that only gives us an indication of what percentage of the students were giving us noise rather than signal, and it is up to us as the analysts to figure out how to make use of that information.

The second thing I would suggest is an

implementation survey to go along with the development and implementation of a new federal program. It doesn't seem like too high a cost to a school that is getting additional money from the federal government to ask them how and to what extent they are implementing the program.

Even though self report isn't always extremely accurate, it is better than no report at all. So having information about whether or not materials are being used as they are planned to be used, or information about whether funds are being used as they were intended to be expended, would be really useful if we wanted to make causal statements about the impact of a federal program. We need to know how the federal program is being implemented and whether the federal program is being implemented as we expected it to be in the various sites. That isn't a small problem.

I have been in a number of school districts that have had new reading adaptations for years before materials get out of the closets in some of the schools. So that is a substantial problem. The nature of the problem is that it tends to bias the results. It is not a matter of accuracy, it is a matter of bias. So you end up with, for the most part, if there is less than decent implementation, with less than optimal effect sizes, because the folks aren't using things the way you expected them to.

The other thing that implementation would give us would be an indicator of the extent to which the program is being implemented. As any good formative evaluator will tell you, that is at least half the battle.

I am going to take a slightly different approach to this cutoff score thing than most people have suggested. I am going to suggest that we try to avoid using cutoff score analyses whenever possible, because to use a technical term, they are really lousy for evaluation procedures.

If you are trying to identify a change that may be relatively subtle, and may be centralized to a particular group of learners, having information that takes a score scale and throws out most of the information by putting kids into finite categories isn't a good way to identify subtle effects, and might be a good way to identify really big effects. But it is unlikely that most of our programs are going to be in the really big effect category, at least given past experience.

It is just including all of these score types that you possibly can in the database. I would further suggest that you recommend to researchers that they avoid using cutoff scores for an evaluation of federal projects whenever possible, because it is really not a good approach. Cutoff scores are useful for a lot of things.

They are useful if you want to identify if a student is going to go on to the next grade. They are useful if you are going to identify whether a person becomes a nurse or a doctor. But they are not overly useful if what you want to do is do an analysis of a program and ask the question, "What is its effect size?"

A test change indicator would be useful. Right now in the database I couldn't figure out any way to see whether either the standards that the state was using, the test instrument that the state was using, or the proficiency levels that the state was using had changed from one year to the next. So that is a pretty nonviolent thing to add, and it would be real useful for an evaluator.

Last point. I would suggest that we consider connections to other data sources. The amount of data that is out there in privately held databases and in state held databases is huge. By limiting ourselves to school level data, we take away a lot of the tools that an evaluator would like to have, like student scores. So the idea of actively seeking out sources of data and connecting the dots.

For instance, let's give an obvious example, the NAEP scores from the states don't seem to be included in the database. So that would be a real obvious one, but there are probably a million others. Since I am talking, I

get to use my favorite example.

One of the things that we do in our spare time at Northwest Evaluation Association is create databases. Right now, we have a growth research database, which has about 35 million longitudinal assessment records, student level data going back as far as 1995. This data isn't on the state assessments, but it is an additional indicator of what the students can do.

I think it is probably important to realize that we don't really need to have information on the state assessments from top to bottom. What we need to know is how the students are doing, how the schools are doing, how the districts are doing, how the states are doing. So connecting the dots to a lot of different data sources is probably a stronger position to stand in.

The data in my particular database are connected to the NCS school identifiers. I think a lot of databases are starting to use that as their common process, because it makes it real easy to go from one year to the next and know how the common core data matches up, know how NAEP data matches up, and know how whatever other data you have of interest matches up. So that is probably an easy way of identifying and maintaining consistency in the schools.

The growth research database, if you think about just that example, could provide researchers with a drill-

down capacity. So if we are looking at the impact of a particular program in a particular state, and we see that there is an effect at the school level, the next question that I think most people would like to ask is, does that school level effect trickle down to be a student level effect. That is something that you can't answer without the right data. So if you connect to, for instance, the growth research database and 47 other databases, you give researchers an opportunity and the ability to decide to do it or not.

So what I would suggest in closing is that whenever there is a question about including data or not including data, include it. It may cost you a little bit more to collect it, but it costs you a lot more if you don't have it, and you decide later on that you need it. So I would suggest erring on the side of inclusion. I would also suggest not minimizing what other folks can bring to the table. The federal government can collect a substantial amount of information, but there are other folks out there that have information, too. So it would be useful to catch what they have and use it as we can.

Thank you.

DR. LUCAS: Don't go away. We have just a few minutes for questions. I have a question. My question is, "What kinds of changes, especially connected to your last

point about how all these other entities have data, as we think about adding the data to the database, what kind of changes do you see might be needed to make sure that the data added has sufficient quality?"

DR. KINGSBURY: That is an excellent question. One of the points that I didn't have a chance to make was that different databases that are out there are kind of variable in terms of the level of quality that they have.

I don't think there has to be a huge amount of work done, because some of it can be left to the evaluators themselves. The other thing I would suggest is that we might not even have to add the data to the database. There are lots of database processes that create live connections to other databases. So one of the things you might want to consider is not taking on the task of making other peoples' data perfect, but allow it its imperfections and require the researchers to identify those. Then if you identify specific problems with specific research databases, then you make those known.

So for instance, the database I was talking about has substantial limitations. It has only those school districts that are working with us. It only has school districts that have gone out of their way to purchase our services, so it is a very self selected group. So that would be something that the researchers would want to take

into account. It is not exactly a quality issue, but it is an issue of self selection that is important to consider.

PARTICIPANT: Just a quick point about the database. You mentioned that it would be good to be able to link to other databases, and I agree wholeheartedly. But the CCD indicator is on there for each record. We put that on there so you could take these data and download them and match up to your schools.

DR. YEN: I just wanted to build off of something that Gage said. Some of the discussion or the unanimous opinion is that score scales or scale scores, or some kind of more continuous scores, are more useful than the proficiency level designation.

One of the things that Gage said is that it is going to be very difficult to find program effects if you are only using the gross categories rather than the score scales. I think this is relevant, because as we hear from the folks who are now in charge of this database, in essence they have made up their mind about what they want to do, and they don't really want to hear a different opinion. They say, in effect, "All I care about is proficiency because that is what No Child Left Behind is about."

But I think in communicating with them, they also care about finding program effects. I don't think anyone

in the federal government wants to know, we spend all this money, and there really are effects out there, but we are not going to be able to find them because we are using the wrong data.

So I think there is a marketing issue here, communicating with them, saying, if you want to find program effects, these other kinds of scores will be more helpful to you. It is not just an issue that a kid doesn't turn from not proficient to proficient overnight. There is a growth that goes on there. These other kinds of scores can help monitor that growth and demonstrate that growth, which is a very positive thing for the nation. I would think that is completely consistent with No Child Left Behind.

So I am looking at Steve, because I don't know who the right person is to do this, but somebody writing a summary of the main ideas coming out of this workshop should be marketing some of the main ideas so that people who are listening to this can understand it, relative to what they care about.

DR. KINGSBURY: Thanks, Wendy. You said it much better than me.

DR. LUCAS: We have time for one more, then we have to go to the next speaker. But we may have some time at the end.

PARTICIPANT: I wonder if you could comment on the possible tradeoff between the comprehensiveness of the database, adding all these other things, and the timeliness of the data, that is, how long it would take to get it put together, and consider the alternative. You put out a core set of data, but you have the linkability, so that as other information is available, different people can link them in.

DR. KINGSBURY: I guess my initial bias, having thought of it for five seconds, would be to tend toward the latter, put out that which we have and create links to those other things that exist and might be useful.

I think, as you know, about nine-tenths of the battle in creating a database structure that is useful is making sure that the data is accurate that is inside it. That is an important issue, but I don't think it is an issue that we want to spend a tremendous amount of time on if we can hold the researchers responsible to do part of it, and if we can enable the research as we find out specific problems with databases.

DR. LUCAS: Now we are going to hear from Dr. Kashka Kubzdela.

DR. KUBZDELA: Hello. My presentation is factual. I am not here to make recommendations on the data system, but perhaps it can tell you about possible capacity

of states to do the kinds of analyses that people are talking about here.

This is a new program. It is called the Statewide Longitudinal Data Assistance Grants Program. It was authorized in 2002 by the Educational Technical Assistance Act, along with the Education Sciences Reform Act. It was funded for the first time in 2005, and there is anticipated money in 2006, about \$24.8 million per year so far.

Eligible applicants are the principal education agencies of the 50 states, the District of Columbia, the Commonwealth of Puerto Rico, the United States Virgin Islands, American Samoa, Guam and the Commonwealth of the Northern Mariana Islands.

The primary goal of the program is that it provides competitive grants to state education agencies to enable them to design, develop and implement statewide longitudinal data systems to efficiently and accurately manage, analyze, disaggregate and use longitudinal individual student data.

The long term goal of this program is to increase the number of states that maintain statewide longitudinal data systems in order to assist them in generating and using accurate and timely data, to meet the reporting requirements at all administrative levels from federal to

local, support data driven decision making at the state, district, school and classroom levels, and facilitate research needed to eliminate achievement gaps and improve learning of all students.

In particular, data systems developed under this program should make it possible for SEAs [state education agencies], LEAs [local education agencies], and researchers to conduct value-added research that utilizes links to longitudinal data on students, teachers, programs, initiatives and interventions, in order to help identify the most cost-effective solutions towards improving instruction and student achievement.

An additional goal of this program is to leverage the work supported with grant funds, to facilitate the design, development, implementation and use of our longitudinal data systems by other state and local education agencies. This will be achieved in part by discriminating lessons learned and non-proprietary products and solutions developed by the recipients of the grant.

In order to qualify for the grants, the states have to meet a number of requirements by the end of the project that is funded by the grant. I will just summarize them here to give you an idea of what will be covered in the end.

States are starting from different points, so

some of these things have already been done by some of the states, and some states are starting from scratch, and thus have a longer way to go. But, by the end of the program, the state must have an enterprise-wide data architecture that includes a data model, data dictionary, business rules and quality assurance procedures. The architecture must be based on analysis of current data systems, plans for future enhancements and analyses of information needs across the SEA and district program offices, schools, classrooms and federal reporting requirements. So this program is really aimed to support the development of data systems that meet a wide variety of needs.

The architecture has to be relational in nature, and allow users to readily link records across time and across information systems. The data types and items to be included in the system must, at a minimum, include all data elements required for NCLB reporting, be maintained in a longitudinal format, allow for meaningful longitudinal analyses of student academic growth and factors affecting it, and preferably also include data elements necessary for research to address the effectiveness of educational programs, development, finances and other central education policy issues.

The state, of course, will have to have a unique permanent statewide student identifier. The data will need

to be vertically integrated, to allow for easy movement of data from local to state to federal levels. Also, there is a maximum participation of all LEAs in the data system required before the project's conclusion.

Effective procedures for protecting the security, confidentiality and integrity of data are also required. Other requirements include procedures for assuring technical quality of data and the longitudinal data system they will feed into, to maximize the validity, reliability and accessibility of statewide cross section and longitudinal data at all administrative levels for evaluation and decision making purposes. A data warehouse or comparable means for managing and storing longitudinally linked data and making it accessible and useful to key stakeholders, especially teachers, schools and districts.

Timely and ongoing exchange of data across institutions within a state, especially between districts, potentially also between secondary and postsecondary institutions, timely and ongoing provision of high quality data reports and ad hoc analyses to teachers, schools, districts and other constituents such as parents, school boards, state and local officials, business community and the general public, and timely, effective and ongoing training of intended users of the intended data system.

The state must also develop clear procedures and

insure timely and ongoing restricted access to data for policy oriented research in conformance with FERPA requirements, and facilitate analyses and rigorous research to evaluate the effectiveness of programs and efficiently improve student learning and academic achievement.

They must develop and implement within the first 15 months of the project clear evaluation criteria for determining successful development and implementation of the data system, and on an ongoing basis evaluate its quality and effectiveness in meeting the reporting and decision support needs of all of its key stakeholders, so that the states find ways to monitor that what they develop is actually useful. Also, they must evaluate the effectiveness of the data system in catalyzing improvement in academic achievement of all students.

The state must also coordinate the use of state and local resources available for educational data systems with the use of federal funds under this grant program, to insure that money from this program supplements and not supplants available state and other funds, and insure that the statewide longitudinal data system developed under this grant will be sustained over time.

Forty-five states applied to this program in its first year; the applications were reviewed by 20 external technical experts who evaluated applications for the

content, quality and feasibility of the data systems proposed. Reviewers also judged the likelihood that each proposed project would have a substantial impact on generating and using accurate and timely data to meet federal, state and local reporting requirements on allowing for value-added and other diagnostic policy relevant research on engaging in data-driven decision making and improving student achievement.

I don't need to go through the process because I am running out of time. [See Dr. Kubzdela's PowerPoint slides for additional information.]

The competition was announced in April of this year. The application was at the end of June. A panel met in September, and awards were announced just a couple of weeks ago in November. Many of the awarded projects began on December 1, and will have to conclude by the end of November 2008, since it is a three-year grant. Grants could be for up to six million dollars over a maximum of three years. As I mentioned, 45 SEAs applied and funding was sufficient to fund the top 14 applications, ranging from \$1.5 to \$5.8 million.

The grants for those interested were awarded to Alaska, Arkansas, California, Connecticut, Florida, Kentucky, Maryland, Michigan, Minnesota, Ohio, Pennsylvania, South Carolina, Tennessee and Wisconsin.

I just quickly wanted to let people know what were the common needs specified by applicants. Some SEAs have developed their system components as an information need arose, and now are recognizing that they don't have some of the basic infrastructure behind other existing systems, and that this is a real barrier to efficiency or continued growth of their data systems. The proposals often included developing an enterprise-wide architecture, comprehensive data dictionaries and otherwise working across programs to build a comprehensive data system for the state.

Integrating existing databases was a big part of many proposals. The SEAs wanted to take apart the data silos that support teacher certification, accountability, student special [education] program participation and others into a single information system. Many states, in addition to links between program, staff and student data, are planning to take fiscal and facilities data to student outcomes.

Access was a common theme. States with existing systems wanted to see them used fully for decision support. Many of the proposals included developing data portals or data warehouses that would be available to district and school staff members as well as the SEA. Many states are incorporating security features into their data systems to

allow researchers access to data without revealing student or staff identity. Many states are developing solid plans to promote research and use the results for decision making.

School districts need training in order to put accurate timely data into the state system, and to use the data that is in a system for improving classroom instruction. It appears that hardware and system development often fails to include support for participant training. This was a common need, and some states have very elaborate staff training plans at all administrative levels, including schools.

There is much variability across districts' student information systems, which often represents a sizeable investment from the district. A number of states are adopting common data standards that build vertical integration, again allowing the data to be moved from local to state and federal.

Among the awarded applications, the common approaches, all states are leveraging existing work to get a dollar value for every dollar of grant money. Integrating different systems was a common theme, but once they proposed to integrate finance, school facilities, student demographic performance and accountability data, they are also including postsecondary data so that it could

measure over time factors associated with student learning.

A number of applicants proposed to expand their current elementary and secondary systems to include postsecondary students, moving from K-12 to K-16 or K-20, and often including preschool and sometimes before. Other states' applications included developing a standard electronic student transcript. This is for the purpose of moving student records between states.

Several SEAs proposed partnerships outside the K-12 system to gain information or expertise. One SEA proposed to work with the state's health and human services agency, another to work with a major research university.

Some examples of innovative ideas. Three-state collaboration to establish a common system building on unique strengths and contributions by each state. That is Wisconsin, Minnesota and Michigan. Providing a guidebook for other states to use in developing longitudinal systems.

Finally, plans for future grants. If there is funding in 2007, we will announce the competition in probably summer, fall 2006, award the next grants probably in early 2007, and we will learn from experience from current grantees in how to revise and shape the application.

So hopefully that will provide some other means of analyzing data.

DR. LUCAS: I think we are out of time. It is time for the next session, so thank you.

**Agenda Item: Session 5: State and District Reactions**

DR. HENRY: For this session we are going to provide a state- and school district-level perspective in reaction to what we have heard so far. I think it would be a good thing to begin by introducing ourselves. We are going to have about three basic phases to this. We are going to go in the order that appears on the program, and each participant will have about five minutes to give an overview. After we have gotten through that cycle, we have some general questions to submit to our panel, and then we will open it up for general audience submission.

Geno, do you want to introduce yourself?

DR. FLORES: Sure. I'm Geno Flores. I am currently a deputy superintendent of the school district in San Diego City Schools, but previous to that, when the organization of this symposium started, I was deputy superintendent for the state of California, in charge of the assessment and accountability branch of operation.

DR. O'REILLY: I'm Joe O'Reilly. I'm from Mesa Schools in Arizona. It is about a 75,000 student district. I am the executive director for student achievement support, which means I oversee research and testing and

community relations. I have also worked on the state technical advisory committee. I have been involved at the U.S. Department of Education's reviews for my AASA on standards and accountability and assessment systems through the No Child Left Behind reviews. So I have had some experience with that also.

DR. CURL: I am Cory Curl. I am in the office of the deputy commissioner. I am in the testing department of education. I do a lot of different work. I am primarily a data consumer in my work, and try to always be there to provide high quality data for all the decision makers that I answer to. I also work with the federal EDEN [Education Data Exchange Network] project, so I will be speaking about that a lot today.

DR. HENRY: I'm Steve Henry. I am general director of research, evaluation and assessment for Topeka Public Schools, where I have been for about 20 years, past president of National Association of Test Directors, frequent peer reviewer for the feds, and also incoming division H vice president for the school evaluation and program development division.

DR. TAYLOR: I'm Robin Taylor. I am the associate secretary of education in Delaware, responsible for assessment and accountability. I am responsible for student, school, district, state and educator

accountability, licensure certification as well as technology.

DR. CHESTER: Mitchell Chester, associate superintendent with the Ohio Department of Education, responsible for policy development, strategic planning and accountability programs.

DR. HENRY: And Robin, we will begin with you.

DR. TAYLOR: I thought I would spend a few minutes talking a little bit about the background of data collection in Delaware. There are four broad points I want to make in my five minutes.

We are that small, cute state that Gary [Miron] referred to yesterday. We have a unique student ID system in Delaware. We have had that since 1984, so individual student data collection is not new to us. We also have a statewide pupil accounting system. It is a software application that is used by most of our districts. It is voluntary; however the state pays for it, so it doesn't cost the districts anything to do data collection if it is used, or to use it. So there is somewhat of an incentive there.

The department maintains a system. We have a common data dictionary across all the districts in the state, and there are certain data elements that are transmitted to a statewide database at certain points in

time, and those points in time are almost on an hourly basis now, so it looks like it is real time.

One of the features of doing it this way is that for example, when a new student shows up at a school, there is a look-up feature, where the school secretary can look up to see if the student already has a unique ID in the state, and if the student does, they can register the student very easily. They just say that the student entered the school on such and such a day. That sort of feature allows for the school where the student exited to be notified right away. The records get transferred electronically, the electronic records and the state testing information gets transferred to that new school immediately. So they do have some kind of information about the student when they enter the new school. So that kind of information is captured and transferred very easily.

We have been at this for about 20 years. We are that small cute state -- I am going to keep reminding you of that -- with a very large, rich data set, that is being expanded. It is K through 12 right now, and we are trying to expand it to PK to 20. So we are trying to get it out through the university system and out into the world of work, so we are working on that.

The four points that I want to make focus a lot

on data integrity and data quality. Data integrity is a really critical piece, whether we are talking about databases at the state level or whether we are talking about them at the federal level. By that, I mean data integrity basically examines the validity by which the inferences can be drawn from the selected data.

Here is an example. This is what I mean. If you are looking at longitudinal data, you have to be aware of the sources of error that are there. For example, you can have the best unique identification system or student level database, but in your student demographics, if the student re-enters each year and that creates a new record each year, and you are trying to look at data that is correct over time, you have got to have certain controls to insure that it is the same student with the same record. So you have got some real considerations there. You have to look at the structure of that student level database. You have to be aware of that, and you have to make sure that you control for certain things.

It does happen that individual student data changes over time, especially things like LEP [Limited English Proficient] status, special ed services and status, those sorts of things. So you have to be able to explain that in any kind of longitudinal database.

The other thing is that business rules or

decision rules, cut scores, those sorts of things, change over time, so you have got to make sure that you have a system that captures the old and the new and can explain the differences between the two.

Data quality, got to have controls in place. States have to do this. We have to insure in our new federal A-133 audits that down to the student record level, data is accurate. So data quality is very important in this.

The last point I want to make, since I only have one minute left, is that the context of all of this is critical. Yes, you can use a scale score to show growth, that is great, but you have got to have context in all of it.

So as we go along, we will make some more points.

PARTICIPANT: We will next hear from Joe O'Reilly.

DR. O'REILLY: When I was first asked to be here, it was being described to me as, "What do districts think about this state level database and data collection?", I thought, that is easy, they are oblivious to it. Even the best informed are probably vaguely aware of it.

My next reaction was to picture a [Gary] Larson cartoon of two deer standing in the forest. One, which we will say is an AIR person or a Westat or maybe a state NAEP

coordinator, puts his arm around the other who has a birthmark which is white and red and white and red in circles and says, "Bummer of a birthmark there, Bob," and it is a target. Many times schools feel like they are a target when we hear that the states or the feds are going to collect data from us. What you call data collection is what we call taking staff and student time. So that is the first reaction of people.

As I listened to the papers, one of the things I thought about was, I experience what is going on day to day in terms of what we are doing at the state level and the school level on getting some of this data in. As I think about the NCLB, it has had a really big impact over the 20 years I have been involved in this, watching how our data has gone. We still have the involuntary non-national test, which is what we mostly talked about there, but there are other things that have gone on.

The states are more alike in terms of the grades tested, who is tested and when. The demographic information is nominally at least to say we are collecting some of the same subgroups. Some of the work previously, like the SIP initiative, resulted in a lot of our student databases having - many times - the same data definitions, so they are SIP compliant, that we school districts buy from various vendors.

The peer review system. Bill [Schafer] has talked about it, but I think that has had a big impact on guiding people. There is an extensive guide which is about 40 pages long, with detailed rubrics as to what is expected, so it is a big document describing what is it that is expected of states. Then the state people are going forward and doing reviews, so they are being brought along and trained as to what is being expected of their system.

So who is tested and how they are tested is becoming more and more alike. Special ed students, ELL [English Language Learners] students, and that sort of thing. Of the 95 percent tested of every subgroup, I can tell you about the 12 Asian students in our high school in the tenth grade who were not tested, and we had to track down from the district why those 12 kids were not tested, because we were at 94.4 percent of our Asians tested. So I now know what happened to every one of our Asians because of the new system. In the past it would have been, "Oh, we lost a few, we got most of the tests in." Now at the school level, it really means something to them that everyone is tested, and that they are tested in the way they should be tested, so that it accounts for accountability.

In the past, in our information systems, people

used to say that the database is out at the schools with the schools. Well, they weren't. The schools didn't care about the databases, they cared about the kids. We in research and some other areas cared about the quality of our data. But now, for the first time, I am seeing the school care a lot more about the quality of the data, because we have kids in different subgroups. So I just wanted to mention that.

Every state has a database that allows you to look at who is there, how they did on the test, who is tested, their performance level and some of the subgroups they are in. So every state has that because they have to report back to the districts.

One of the things that was mentioned was context. Context is very important, because schools are focused on kids and schools and achievement, and not on research. I have seen superintendents say things like, "you guys get the treatment and don't worry, I'll take care of you." You have got to do something for your kids, too.

The bubble kids that Laura [Hamilton] mentioned, that is something we are having arguments on all the time, or discussions about. We are doing an implementation now with a reading program, and it is "who do you treat, and who do you treat first?" If we have a system where they are looking at the percent passing the standard, they are

going to treat very different things than if we have a gain system. So we have been going back and forth with the districts.

DR. HENRY: You've got one minute.

DR. O'REILLY: One minute, okay. The importance of addressing the issues, because if we don't do it, someone else will do it. Just for the Kids has come to Arizona; it's got all kinds of student level data. They have gotten money from the state to do it, because they promised some things back. The commercial researchers, I won't tell the Gage story, but he made a comparison between his company and Gage's company because they know about marketing and all that, but Gage, they are just interested in creating good tests, not making money the way the other people are.

We already addressed the other thing, which I think hadn't been addressed as well until Wendy Yen addressed it with her disconnect between what we heard the other day, that we are just going to have performance levels, and what we have been talking about for two days is all the other things that are needed. I think Wendy said that very well.

DR. HENRY: Now we will hear from Cory Curl.

DR. CURL: I'm going to speak from the state perspective. I have been PBDMI [Performance-Based Data

Management Initiative] coordinator in Tennessee for two years now. So when we are talking about using school level data to evaluate federal programs, the EDEN data set is what is most familiar to me, so I will be speaking a great deal about that.

One thing I would like to make sure to comment on to the U.S. Department of Education with PBDMI is that they involve states in this process from the very beginning. When PBDMI started, they went around to every single state education agency and asked us very specifically about what data elements we collect and when and who is responsible for those.

As they began to create EDEN, which is the database in the submissions system, they always looked to states as partners. They always spoke to us as if we were partners in that process. So even though the U.S. Department was asking us for two years, two to three years, to be duplicating all of our data reporting, you need to send things electronically for EDEN, and we need you to still send in your paper reports, I think a lot of states really bought into this process, because of the fact that we were brought into it from the very beginning.

The other thing is that EDEN to us makes sense. This is data that once it hits the federal level will actually be useful for some basic program evaluation.

What we have been doing, which is sending in long Microsoft Word documents with lots of narrative in there, with data that was typed in by various people, once we send that off and it goes to the federal level, we don't really know what it is going to be used for, if at all. Now we know that this data is going to be used.

The data that is there is very specific. I think it is very specific to what the goals of NCLB are. I think that is why they are asking us for the achievement levels.

It is also very much mirrored in our other federal reporting requirements. The same data that we send to EDEN is also the same data that we are required to put on our state report cards, so it is a fairly efficient process for us.

What I envision would happen once we submit this data to the federal level is that then it starts to get into some program officer's thought processes, and it spurs more questions that then they begin to take to develop the more formal program evaluations that go on.

The data elements in EDEN are extensive. It goes from AYP [adequate yearly progress] assessment data, dropout data, graduation data, all by subgroups, financial data on federal programs, which I think is very useful, participation data in federal programs, from Title 1 neglected, delinquent, migrant, a lot of data on special

ed, a lot of data on teachers.

Because this is data that we are familiar with, it is fairly easy for us to report it. I do believe that in a few years, you are going to see fairly universal participation in this process, especially if the Secretary decides to make it mandatory.

But there is a lot about EDEN that is not easy for us, because we have been used to sending in paper reports. A lot of the data that we are used to submitting is at the state level. So the fact of drilling down to a district or a school level is very innovative for us. In fact, in my state we won't be submitting any -- unless a miracle happens -- any federal programs participation data except at the state level, because it is just very difficult for us to collect it.

So what we are trying to do is go forward and decide what is the best way for us to begin collecting this data. We believe it is at the student level. So that means that we will begin to utilize the grant that we received to start building in the processes so we are collecting this data correctly.

One really good thing about EDEN is that it really maximizes the data quality. When we submit it electronically, we are not just blindly sending it off. It is going through some business rules, so if it is not right

it gets kicked back to us and so we fix it. The definitions are clear across all the states, and there is, I hope, going to be a really fantastic meta-data repository that is going to be fascinating, not only for data analysis, but for policy analysis on differences across the states.

I imagine that EDEN will evolve a great deal over time. I think the discussions that we are having here are very valuable. What can we do to EDEN that isn't going to be difficult for states, that can still make it more useful from a research-oriented perspective. I also think that as states begin to use more longitudinal growth measures over time, that there could be some school-level, district-level and growth- or value-added type statistics that might be valuable to add in the future.

One thing that the U.S. Department is willing to do is to make EDEN available to researchers and to others, which I don't think has happened yet.

What I would say to the research community here is, states are at a pretty revolutionary time right now. We have some federal money that we have never had before. We have a lot of pressure on us from our schools and districts that want to have really sophisticated data available to them.

We are making a lot of changes. We have got

proposals on the table every day saying which data are we going to collect, how are we going to collect this data. I think this is a great time for researchers to get involved with your state programs and say, if you just collect this one element this particular way, it could allow us to answer some big questions.

So that is my suggestions to you.

DR. HENRY: Now we are ready for Mitch.

DR. CHESTER: I'll try to speak up. I am going to address the issue of research questions, I think primarily in my comments, and I hope get a little bit to the potential for longitudinal student-level data.

I want to make three points in regard to research questions. I hope that these are useful, given the topic of using school-level data for evaluating federal programs. I suspect that what I am going to say goes well beyond federal programs.

The first point that I want to make is that the questions that are often of interest are not so much [comment not caught by microphone] and just to use the example that Elizabeth [Stuart] used, "Is investment in libraries helping literacy?", but the questions are usually more nuanced that we are interested in at the state level: "Under what conditions, or given this context, is this investment likely to pay off?". It is not so much, "Does

it work or doesn't it work?"

Charter schools is another great example of that. Charter schools are not a universal phenomenon. The question isn't so much are they successful or aren't they successful. That certainly was the question when they were starting up, but at this point in time it is, what do we know about the situations where they are successful, what do we know about the situations where they are not successful.

This is critical around school improvement efforts at the state level. Given our ability to size up district capacity, school capacity, what kind of treatment is likely to be successful? So the question isn't, "Is treatment A successful?" The question that really would be helpful is, "Under what conditions, what circumstances?"

The second point about research questions. I don't think I heard a lot about this, it may have been touched on, but there was a discussion in day one about the notion of counter-factual, and what is the correct comparison group, what is the basis on which the conclusion of effectiveness is met.

I would propose that a key counter-factual for states right now is whether or not a treatment results in a school or a district meeting a target. Certainly AYP is the federal target. In some states the target is also a

rating, how we classify schools into five rating categories. So we are really interested in, given a particular treatment intervention, what is the likelihood - the counter-factual isn't so much did that school do better than comparable schools. A key issue is, did that treatment, did that intervention, was it sufficient to move that school above the target. If it wasn't, the fact that it helped that school perform a little better than comparable schools doesn't quite get us all the way there.

The third piece of research questions is helping unpack policy attributes that define effective state policies. We have been doing a lot of thinking in Ohio to try and be more explicit about a theory of action, about what is it we do as a state agency that has a payoff down the line that has an influence on student achievement. When we look around at the research in this area, it is pretty thin. There is not a lot to draw on here.

There is general agreement, and that I think there is general buy-in, to the standards-based reform notion and the notion of alignment, clarifying what you want kids to learn, measuring against it, holding people accountable, developing supports to get there, the whole notion that the more that system is aligned, the better the results are going to be. But there are an awful lot of black boxes along the way, and for a state agency in

particular, there are a lot of questions about what is the state's role in that, how does the state best leverage that happening and so forth. So three notions around research questions.

I want to say one thing about context and policy context. Again, this circles back to my first comment. There have just got to be these huge interactions between these treatments and these interventions in the various policy contexts in which they are employed. If you are doing research in Ohio right now, and you are comparing impact in Dayton to impact in Columbus, and you are not aware of what has happened in Dayton over the last decade around community schools and the loss of the student population and who is left in the school system, you are kind of flying blindly. So I think context is important, and certainly varies from state to state. So having information in the database about a variation in a policy context, to the extent to which a state is embracing choices, is going toward vouchers and so forth, I think is important to understanding the phenomenon.

I'll close with a comment about longitudinal student-level data. One of the things that NCLB has done is, it has provided an opportunity for states, including Ohio, who weren't testing every grade three through eight, now are testing three through eight. And given the fact

that we now have this data set, we are very interested in linking it and making sure that we have longitudinal records, and taking advantage of the data that that testing regime brings to us.

So we are creating a statewide database, student level IDs and so forth, and I think more and more states are. There are states like Delaware that have been at it for a long time, well before NCLB.

We have also secured legislation, because this had major political ramifications, to get a state ID in place. When it was first put in place through legislation, it was very restricted in its use. In fact, it was of little value to us, because it was so restricted in the legislation.

So I suspect there is opportunity here, and it probably varies from one state to another for the research community, the U.S. Department of Ed, to start to think about how to take advantage of those states that are creating these databases.

Thank you.

DR. HENRY: Now we have got a few prepared questions to lob out here to our group, and you can have it, whoever wants to respond, all of you or selectively.

Here is the first question. A little context. We know that the databases originally put together by AIR

was envisioned for federal application. In particular, there was a lot of interest in the potential for application for federal program evaluations, which was a central piece of why we are here.

But given that, what sort of applications do you envision that might be possible at the state level or district level? Anybody have any thoughts about that?

PARTICIPANT: One of the things that we talked about, or that came up yesterday, Judith talked about an inventory of interventions and what worked and what didn't. Oftentimes, principals are saying, we are working as hard as we can, we are trying everything we can, what do I do? They come asking for advice, what works. If we can tie some of this and have some studies done on what are the interventions, what works with different populations, which also argues for student level data, if we are interested in limiting those proficient students specifically -- I'd rather look at limited English proficient students than schools that have large numbers of LEP students -- so providing schools with information on what works and doing research on what works with a database like this would be very helpful.

It would be much better than the publisher who comes in and does a nice little study. I saw in that week, our kids increased 18 percent more than the control group,

but they learned less than one more answer right on this post test, which I know because I know what the test is. All that comes out is, they grow way more with their product than with the non-product.

So I would much rather rely on something like this than some of the other data that we get out there.

DR. TAYLOR: I think I would add, Mitch is right on about the questions. To us at the state level, the important thing is the questions that you want to ask as researchers. We welcome those questions. We will give you the data to do that.

The big issue that we have is, most of the states are now in the situation where we are building student-level databases. How much effort do we really want to do to duplicate this at the federal level? It is a duplication. To what extent do we want to spend money to duplicate something that states already have?

My question back to everybody would be, is there a way to take what states are already doing, and take the individual databases that states are already doing, and hook into that, do something with that, so that we have some sort of common context for some of the data elements, not all of them, but some of the data elements?

The other comment that I would make about that is that you have to understand that coming from a state that

has a rich data set, we do still have a fear. It is not a fear of researchers, it is a fear of reporting, and it is a fear of how some of the data is going to get reported by some third party entities, because they are still out there ranking and talking about one district that out-performs the other district, and they are making those kinds of judgments, which gets totally away from program effectiveness, and it leave us gun shy.

PARTICIPANT: [Initial comments not caught by microphone] -- and hence, here is an area we could do with some extra help, here is an area where you want to read what we have got before you propose something. I think your nervousness about the lead table approach is something which evaluators are now very worried about. But you have really got to put that in capital letters at the bottom. It is not enough for them to think that releasing research reports is the end of the story. The end of the story may be the end of you, not that story.

PARTICIPANT: If I might while the mike is shifting over there, when I was director in Maryland, we echoed the same concern that you have about researchers and having gotten burned a few times, the discrimination going on ahead of our opportunity to comment on it.

So whenever we entered into an agreement with a researcher, we decided to write in the need for not a

censoring, but an opportunity to review and comment before publication.

DR. TAYLOR: I'll be a little bit clearer. I'm not talking about researchers, the kind of researchers we have got in the room today. I'm talking about things that are happening to states on websites, where we give data. I will be blunt, I am talking about Standard & Poor's and the metrics that they use with your data.

So there is a gun shyness on our part that doesn't have anything to do with real research. It has something to do with taking data and using it in a certain way.

PARTICIPANT: Well, there is a tension here. From the state perspective, certainly there is a need to at the least not tear down the efforts that we are making, so we are kind of defensive on that.

But the tension on the other side is -- and this is to Stephen's question about how this database can be helpful -- I think all of us in one way or another, we may not look like it, are interested in learning from cross-state comparisons. A lot of what we can learn by simply sticking with Ohio's data is limited. The cross-state comparisons may say to us, we should be looking at these three other states, because they seem to be getting better results than we are, or they seem to be closing the

achievement gap at a rate that we are not succeeding, and we better find out what is going on in those states.

PARTICIPANT: I'll speak a little bit to the benefits of a more broad database. Many federal programs, the outcome that they are looking for isn't just growth in student achievement. A lot of times it is also educational attainment measures and the ability to look at dropout rates, graduation rates, across states and have the data to provide some context for those I think is extremely valuable.

I should probably say also that one of the major data policy issues going on in the nation right now is a major focus on how to make dropout and graduation data of better understand, so that those cross state comparisons can have greater validity.

DR. FLORES: Steve and I are hosts, and we are supposed to hold back and let our invited guests speak, but I would just like to chime in with Robin's comment. I guess now that I have left the state, maybe the headache won't be there as greatly, but I certainly was chasing our good friends at the Harvard Civil Rights Group all across the state of California, because I kept saying everywhere I went, that is not a dropout report, it is an accumulated promotion index, and helping somebody to understand the difference between the two takes too much time. The

newspaper people say, you know what? Never mind, I'm going to use the dropout rate, anyway.

So sometimes the story gets out ahead of you before you get a chance to help the public understand exactly what is being reported.

The only other piece, and then I will leave it back to our guests to speak, having been at the state level, getting common core data elements together is really like birthing lots of elephants, especially in a large bureaucracy, because everybody has their own little silo of data by various programs. A simple and yet crazy example that I always gave was that in the California department of education we collected data four different ways on sex. It was boy, girl, male, female, A, B, one, two and other. There were lots of data elements in the other column. Nobody wanted to give it all up just to be able to get to a common way of collecting the reporting out. It takes some major efforts out of state level and then to then implement those across a large state or any state was very difficult.

I'll turn it back over to our guests.

PARTICIPANT: I was thinking about this from my own experience too, to toss something in. I have an example that isn't literally from this database structure that we have been contemplating, but it deals with a subpart of it at our state level.

Different elements of the database content that we have talked about exist in various places. Our state made available to us a statewide data set that had every district, every school in the state, many of the variables that are part of the other set. One practical application that we made of that in our district was to just go through the state files and find schools in the state that had high concentrations of minority and low income students and yet, despite that produced high performing, high scores for their buildings. Then we contacted those schools and districts to learn a little bit more about their practices.

So the more the better for this kind of data set that makes inquiry based things like that. It doesn't have to be a full-blown high-level research thing. As other people said, description is good, and supporting inquiry and exploration is also good.

DR. HENRY: Another question that we have to chew on here a little bit is one that was touched on yesterday a little bit, the unit of analysis, the level of analysis. What do you all think is the right unit of analysis? This database as currently configured seems to be oriented toward school-level, that is pretty clear. What sort of fit is that? How do you feel about it from a state and a district perspective? Is that an appropriate unit to make the unit of study or not? Or would other things be good,

too? Comments?

PARTICIPANT: I guess I don't have anything new to say on this topic. It seems like the more you can have, the better. It seems to me that one thing that NCLB has shined a light on is the fact that when you look at aggregate school data, you are not necessarily seeing the whole story. You may see a generally high-performing school, but there may be groups of kids within that school who aren't enjoying that level of success.

So I would like to eat my cake and have it, too. I'd like to have it from the student-level on up. I think that part of the story typically is, the more you can drill down, the better you understand the story. I have paid more and more attention to school finance and fiscal management issues. When your state data system, as ours is, is only robust at the district level and not at the school level, you are probably missing much of the story, because much of the story in school finance is probably below the district level. It is how funds are allocated at the school level.

PARTICIPANT: Again, I would repeat a little bit of what I have said. I have no problem with student level data, because I believe that is where you are going to be able to answer your questions, and that is the only way you are going to really be able to thoroughly answer your

questions.

So I think unit of analysis, clearly student-level is the best. But I would go back and I would ask the very same question that I asked just a minute ago. I would reframe it, though. How can you do this? How can this be done in such a way that it doesn't duplicate what states are already doing? Or is there a way that we can build on what states are already doing, so that we can minimize the cost and minimize the human expense of doing all of this.

Creating a national student level database is going to be extremely expensive. But I do believe that we need to work together to create common data elements, data elements with a common definition across the states. That would help all of us be able to do a better job of what we are doing.

So I think there are a variety of ways that we could think about this. We don't necessarily have to have one repository. Maybe there is some way we can have common data elements across the states. Maybe states can develop a student level database that everybody can tap into. There are certain data elements that are extracted, or that are available for everybody to get, and there is some way to insure privacy and security and all those kinds of things.

So I believe the unit of analysis should be down

at the student level. That is where the best information is going to be, but we need to be smart about how to do it.

PARTICIPANT: So many options. I would say that at this point, the goal of building state capacity to manage this data is a major priority. Making sure that the data that states are collecting is accurate and is the data that we need to answer our questions is a top priority, and providing access to that data to answer the questions is a requirement. But building that state capacity and making sure that that data is accurate for research is I think the concern, and where the data is housed or who owns it I think is secondary.

PARTICIPANT: There are a number of questions that we would want student-level data for, but oftentimes our interventions are at the school level, so oftentimes school-level data is enough.

The challenges that are presented by the student level -- I'm not as up to speed on this, it wasn't really made explicit, but it was like, "we can't go there," so I assume there is some reason why the Department can't get student-level data. But at least we can get it from some places where they have good data. If we can at least start to get some of that data together and see what are the challenges to analyze it, maybe verify that we can use school level data, look at school level data and look at

the same question with student level data, and make sure some of your assumptions would be verified.

So I think there are some very valuable reasons. I would prefer student level data overall, but I think there are sometimes when school level data would be adequate. We should at least make steps in the right direction, even if we can't get the whole way with every state having good student data and having the statistics that could handle the student database from every student in the country.

DR. HENRY: Anyone else?

PARTICIPANT: To somebody listening who ran the only federal program on teacher evaluation, there is a pretty obvious missing link here. What about classroom level? I know it wreaks havoc with the union situation, but there are all sorts of ways; we won't have to threaten teachers in a direct way.

But compared with NCLB and anything else we are talking about, this is all peanuts compared with the intervention of the teacher. Yet, we are not addressing the middle level of the aggregation problem at the classroom level. This could be so helpful, even if we made slow movement towards that, perhaps with the elementary schools where the teacher is connected to the individual students in a very much more tight way than in the high

school, we could begin to get some information.

I remember, Cory, in the early days in Tennessee, when Bill [Sanders] was running the pilot studies, he discovered these two teachers in this one school who had managed to produce an actual negative gain across a year of teaching. He asked the principal about this, and the principal said, "You have to be kidding, these are two of my oldest and most reliable teachers." Bill said, "Take a week and talk to people, and then come back and talk to me again." A week later, the guy came back and said, "Every parent here knows to move their kids out of these classrooms the moment they are put in them. Every other teacher knows this, but I didn't know it. Well, I learned something."

Well, that is a lesson that we could learn with the data that we have already in a number of states actually be made possible for the principal to use. It is a shame if we don't also address that in the course of our remarks here.

DR. TAYLOR: I want to respond to that just briefly. For the first time this year, we are collecting data at the teacher level. We know what teacher is teaching what students, so we have that information, and we are trying to tie the percent of students being taught by a not highly qualified teacher, we are trying to tie that

down so that we can answer that at the classroom level from the state.

The unique thing in Delaware is that our schools don't report. The state reports on behalf of all of our schools. So we are trying to get that data. We have kind of crossed the line with the union.

The other thing that we are doing is, we have a statewide educator appraisal system in Delaware, where all teachers and building level administrators are appraised the same way. We have had that since the mid-80s. We are doing version two of that. A very controversial piece of that is that 20 percent of that is based on classroom assessment data.

There are two parts to that. One is the teacher-made test and the other is the state assessment down at the classroom level. The question that the appraisers are asking teachers is, "How much did your students improve, and how can you show me that using the data that you have?" That data is coming out of the state database.

So it is very interesting. That is being piloted in two districts. So that is out.

PARTICIPANT: I just wanted to say something about your plea to somehow leverage the data that exists and being built in states. I am taking it back to you. The National Center for Education Statistics has a forum

which is trying to work very hard to provide data dictionaries and other forms. If they are not sufficient, if they are not strong enough, if they are not doing that job, then I think if the states voice that, rather than have it be something that comes down from the federal government, which is not my decision, but if the states voice a desire for it, I think that would be an answer, and things could be taken to whatever level would be helpful to everybody.

DR. HENRY: We have crossed into our time period for questions, if there are any other questions from the field that you would like to ask of our panel members.

PARTICIPANT: A question regarding SIF. To what degree is SIF incorporated into your data systems, particularly relating to curriculum and assessment? And how much has SIF and the Department's support of SIF been a factor in affecting your data collection systems and data management and any other factors?

PARTICIPANT: We are beginning to explore the option of collecting all of this data that we are talking about at the student level that we haven't been collecting at the student level through a pilot with some districts using SIF.

When we created our student level enrollment tracking system, it was before NCLB, so there is a great

deal of additional student level data elements that we would like to collect so that our SIF is compatible. So we will be working with some districts. They will be able to leverage the horizontal integration capabilities of SIF, and then we will be interfacing with them for vertical reporting. Then as we learn from that pilot, we have to take it statewide.

PARTICIPANT: I would just add that we too developed most of our student level database information prior to SIF. However, we are trying very hard to make everything SIF compliant as we move forward. That is one of our goals in the state with our management systems, all of our management systems, actually.

DR. O'REILLY: Arizona has developed their student information system more recently, so it is SIF compliant. It caused changes in the districts that hadn't changed over their student information system to a more recent system that already was set up that way.

PARTICIPANT: It looks like the late afternoon of the second day of the ending session.

DR. HENRY: We did have another question that we talked about. In fact, in June, when our steering committee got together and we brainstormed the whole range of topics to be considered for this session, one of the areas that we got into was data quality. As it turns out,

that didn't formally get on the agenda, although several people have brought it up.

But let's kick that around a little bit. What thoughts do you have about data quality as it relates to this database that you might be able to relate to your state and district perspectives?

PARTICIPANT: I just want to say, I'm in favor of data quality.

DR. TAYLOR: Me, too. I'll be the me-too on that. I'm the one that usually has the -- when Mitch and I go to meetings, I'm the one that usually has the data quality stories to tell. I do have one that I will share with you.

I have been doing this for a long, long time, but there is a new element to data quality that you have to look at now, and that is the games that the schools and the districts are playing with data elements. It is not an issue of, are they reporting; they certainly are, but let me give you a very clear example.

This occurred to me last fall. It was through our electronic audits that I caught this. One of the phenomena that occurred last year is, we were getting ready to send the test vendor our electronic student identification system, as to who was going to be taking the ninth grade test, who was going to be taking the tenth

grade test. One of the things that I noticed is that there were some high schools that promoted one-third to one-half of their ninth graders to grade 11, in the middle of the school year.

Well, guess what? They earned the Carnegie units to be 11th graders, that is probably very true. They were enrolled as ninth graders in September, come February they are now 11th graders. They skipped the tenth grade test, which was the test that was being used for school accountability. It was legitimate, and they had done nothing wrong.

But that brings up some very serious policy issues, and it is a data quality issue as well. The outcomes that are being used to judge schools aren't going to be totally accurate because of some games that people were playing.

So I understand where people are coming from and what people are doing right now to game them.

DR. CHESTER: I kid Robin. She thinks that is an example of lack of data integrity. I say that is an example of a state bureaucrat getting in the way of innovation, which does happen with these data systems. I keep getting that complaint, that it forces people to put kids in slots that they prefer not to put them into.

Here is an experience that Ohio had. I don't

know how generalizable it is, but we instituted a statewide unique student identifier [SSID] about three years ago. One of the outcomes of that was that in the short run, we uncovered lots more issues with data problems than we had ever seen before. A lot of it had fiscal impacts for the state, kids being double counted and counted in a charter school and in the regular school.

So the implementation of the SSID let us be a lot more precise about auditing data and validating data than we had ever been in the past. The short-range outcome of that was, it looked like we had a lot more problem problems than we had ever had.

The reality is that we probably didn't have any more data problems; it is just that we now had the ability to find them.

PARTICIPANT: I think we are running into some similar issues. I think a lot of states have traditionally collected a lot of their data at the aggregate level. So we were counting on districts and schools to make a lot of decisions about what that data looks like.

For example, now we are moving to collecting data at the student level. So when you look at our dropout data at the individual student level, you will find many instances where students drop out of school several times during the year from different schools. Sometimes it is

within the same district and sometimes it is not. So when you are calculating counts of students who dropped out, to whom do you assign that dropout? Is it the first high school, the second, all of them? And what do you do when you aggregate it up to the district level and the state level?

So those are a lot of decisions that drive our data quality that we haven't had to make before.

PARTICIPANT: I just want to make a quick comment on Cory's point. It has to do with transparency and the business rules and the decision rules. One of the things that has to happen, the more states build the student-level database systems, there has got to be technical and operational manuals that go along with the student database systems, that clearly define the decisions, the business rules, the data elements, and includes a piece on what quality assurances or what quality control is being done on the data. So I just throw that out as well.

PARTICIPANT: Certainly what we are doing now is much better than the self report that we used to have, whether or not they were English language learners. How many kids figured, I'm learning English? So that was a real problem.

It is much better than it used to be. In Arizona they have actually used the student database also for

funding. So if you don't have the kids in the right program, you are not going to get funded for that kid for ELL or for special ed. So there are other reasons why the data need to be accurate that people care a lot more about than research or even accountability purposes. So I think that helps a little bit, how the whole system is being used and not just for research.

DR. FLORES: Let me just add a couple of quick ones. Not to be narrative with examples of poor quality, but one of the issues I discovered from the state level, now at a large district, is that over time, the traditional enrollment process of walking to your neighborhood school and filling out the papers and getting enrolled in school, because of the requirements of so many more data elements in there, it is only as good as that clerk at that school. So you either have a massive training and professional development requirement, or you have got to create an enrollment center someplace else. I can't count on that secretary at the school making sure that they have entered everything properly.

Similarly, the other thing that happens very commonly, especially around the start of the year, you would love to have parents and children come and enroll a week before school or two weeks or something, with plenty of time, but often they show up on the first or second day,

because they see the buses going down, and it must be time for school.

Well, schools are loving, caring places, and those clerks there are also loving, caring people. They don't want the parents sitting in the office, they want the child in the class. So they will take down a name and maybe an address and a few other pieces of information, and send them on their way, always with good intentions that they will go back and fill in the data later, and later never comes, until somebody from the state agency says, we are missing some key pieces.

The only other [point] that I would add in, for those of you who have much greater authority in recommendations to the federal government is, give systems time to get it, whatever the it is that you are going to want. To suddenly ask for something now, this year, which sounds great, if you are going to ask for it in actual two years from now, that is when you want to use it. But suddenly trying to go out and get something is awfully difficult.

An interesting issue for example would be the highly qualified teacher. If you are going to want to know if they did it by examination or if they did it by the state house procedures, this is an important piece to know if that is going to be an important element in your

evaluation reporting. If so, that is a whole different type of collection; are you qualified or are you NCLB compliant, yes, no, check the box.

DR. HENRY: We have time for probably one more question.

PARTICIPANT: Do you find that when you share research files with researchers, you hear back from them about quality issues that they have run into?

DR. O'REILLY: We are lucky if we hear back as to what the results are. So, no.

DR. TAYLOR: I would add that sometimes we do. Sometimes they will call and they will say, what does this mean, they will ask the context. So I would say that yes, we do hear back.

But Joe is right. Unless somebody is calling you and asking you for the data, you don't always know what happens in the end. But for the most part, most researchers that want our data call and ask what it means, and that is a good thing. Keep it up. That is what we like.

PARTICIPANT: I wondered if you all could comment on something that has been talked about quite a bit over the past two days. We have heard a lot about how difficult it is to do any analysis on percent proficient type of outcome data, if that was the only kind of data in the

database. We have heard that one of the reasons for moving to that is that that is what is required out of NCLB.

So I wondered if you could comment on whether you actually maintain more than percent proficient, and at what levels you maintain it. Do you have means and variance, means and standard deviations, percentiles? Do you calculate them only at the state level or at the building level? I just wondered if you all would comment on that a bit.

PARTICIPANT: We certainly have a lot more than just the percent of students following into each of the performance levels, right down to the student level. We have all of that data, no question about that.

PARTICIPANT: I would echo that. We do a lot more than just collect percent proficient at the school level. We report a lot more than just percent proficient at the school level in our school report cards and things like that. So the data is there, there is data.

I would also add that I very much agree with you, and I very much think that you are on the right track. Percent proficient tells one story, but the scale score and the standard deviation, all that kind of information, is critical to tell the whole story. You have got to look at the big picture here, you can't just focus in on one little piece of it. So I would very much agree that other

information is critical to this.

PARTICIPANT: Even at the school level, we are looking at sub-scales, not just scale scores. In fact, they don't look at scale score so much; it is performance level and sub-scales, where students -- where are they, where are they strong, where are they weak, and how close are they to meeting the standard or approaching the standard.

In fact, the grade levels all get together with their kids coming in who have not mastered the previous grade standards, and their kids are listed from closest to pass and the farthest. What they focus on are the sub-scales, where kids have common deficits that we need to work on. So it gets down to that level of specificity at the classroom level, then at the school level there are averages, and at the district, and it goes on up.

DR. FLORES: Coming from California, I would also add that the interesting piece, the public relations approach as well, all that you see in the paper and all of the reports always describe percent proficient. We are trying to build an understanding of key policy makers, whether they be state-appointed board members or others, about other ways of reporting change over time, rather than just percent proficient.

But many schools and policy people are stuck in

that realm, because that is all they read, or that is what gets quick press. Sometimes they might think there is some deviance behind wanting to report it differently, as if people will take their eye off the prize of getting proficient and be satisfied that we are being able to see some change or some growth or some improvement, and we will have to quickly counter by saying, that is how we know they are getting ready to be proficient. You have to try to explain to them, otherwise they think that we are trying to hide information.

PARTICIPANT: I would add one other thing. In our accountability system, our other indicator for elementary and middle schools is the change in the percent of students not meeting the standard. That is looked at by scale scores. So that is what we look at for our other indicator in elementary and middle schools in our accountability system.

PARTICIPANT: Somebody who has gone to districts and states for data, I have always been met with very good cooperation. The only fault there is in the collaboration is usually on the part of the researcher. And of course, you are under a time constraint. The researcher goes back, gets the data, so happy, he forgets about what he is trying to do for two months, and then go back to you. So you are not at fault for not being able to respond immediately.

Amplifying on the point about using the performance levels, it would be graded data. There is only so much you can do when you are trying to detect changes. When we started using multilevel analysis and a new study would come in, so we can handle categorical outcomes, and people are very excited because there are a lot of categorical data, and let's take a look and see if it is a time series of zero and one, and can we estimate growth and then talk about the reliability of the growth that we are observing.

There is not enough information, given the number of time points you have, to give you a good estimate of the slopes. That ease of variability is not available for you to study across individuals or across schools. That is one major impact of using the scale that has the graded information. Given the same number of time points you have, you just have to log that data. A long time series, I think you can do something with it, but a short time series, you can't.

DR. FLORES: What I would request from a state is, if that is the kind of data element that you would like to have added to your database, that you do work with states and districts to be thoughtful about exactly what that definition of the data element would be, and then give us enough time to implement it. Then we can report

whatever.

DR. HENRY: We have used up our time. I would like to thank our panelists for their contributions. We are ready for a break.

DR. FLORES: There is a break of about 15 minutes, and then we will start the next session right at 3:30.

(Brief recess.)

**Agenda Item: Session 6: Larger Contexts for the Database**

DR. YEN: Given the limited time we have, we are very lucky to have Don McLaughlin here, who has done so much work on this database, and who has lots of information to share with us. So we are just going to charge right into that. Take it away, Don.

DR. MC LAUGHLIN: First, I would like to thank the NRC for inviting such smart and eloquent people here to discuss this database and issues about it that I have been wrestling with for much of the last ten years. I would like to thank the Department of Education for funding NRC for this panel. I would like to thank all those people representing states who over the years have provided data for this database.

I have more than a half an hour's worth to talk about, so I am going to move quickly, and I am going to

skip some of the slides.

The outline we can skip over. The history I talked about yesterday morning, so I can skip over that. I was going to talk a bit about inferring causality, but I am in the me-too mode on that. I think that was discussed quite well over the last couple of days. I certainly think there are a great deal of problems with trying to make causal inferences, and with the kinds of data that we have available in the database, we need a lot more things like random control.

What I want to talk about, and I have five different topics, are some of the things I have learned about the database, and some of the constraints and how we can deal with them. I learned the idea of trying to come up with an approximate answer to the right question, questions that people really want to know the answer for, even if you can only come up with an approximate question from those hours in the Gauss house with John Tukey back in the '60s. I also learned from him the importance of looking at data, not just theories.

So to discuss some of these issues, I figured I would look at the data because of things like school versus student. On the one hand, that is an issue of what research question you are addressing, but there is also the question of how much difference does it make. For the

question about the different measures that you have in the different states, there is the question of how much difference it makes. You may be not matching the assumption perfectly, but you may be matching it approximately. So let me go ahead.

The first issue is, the school is the unit of analysis. How different are student population achievement statistics based on school averages from those based directly on student records? I can move quickly, because I think that you all know about means and standard deviations.

The database I am using for this, if we are going to compare school and student records, you need student records that you can aggregate to the school level. What I am reporting on is another four-state study that we did in 2003 in four states. We got the student state assessment records for all of the NAEP participants in fourth and eighth grade reading for those four states, I think 40,000 students total. They were matched.

Then one of the first steps in doing an analysis like this is to standardize the measures across the state, so I did this for this analysis, for this report, to a mean of 250 in each of the four states and a standard deviation of 50 at the student level. What would the means and standard deviations be at the school level? They aren't

affected. The standard deviations are 35 to 50 percent as large.

The surprising thing, if you look at the paper that was in the folders, across the four states these statistics are worth noting, because there isn't a great deal of variation between states and between the assessments. You notice the standard deviations are quite similar in reading and in math and in fourth and eighth grade. So this tells us something about the nature of the system.

I like to look at population profiles. The population profile is the graph of the achievement measure. This happens to be the NAEP achievement score by the percentile of the population. You can do that where you graph for each student the student score, or where you graph the school mean of the student. What we see is that the means are the same. The variance for the school level profile is flatter, it has less variance, plus they have the same shape. In fact, on both of these, if you get out a ruler and measure it, the tail-off on the left end is a larger tail-off. That is, it is a scanned distribution, the same kind of scan that shows up for the school and student profiles. The distance between the first and 20th percentile is larger than the distance between the 80th and the 100th percentile.

How about relations? That is what we are more interested in. I took for this session one example. In the four states I did regressions, predicting from free lunch eligibility the achievement scores and record the standard score difference associated with all versus none of the students.

Now, at the student level this means zero or one, where the student is or is not free lunch eligible on the record. We find that the regression coefficient is, if you pick two students and one is free lunch eligible and the other isn't within each of these four states, you would find that their scores on average tend to be about three-quarters of the standard deviation lower than those without.

If we first take school means and then do the same regression in school means, if everything were random, there is no school effect. With the regression coefficients, you would tend to expect to get the same result. In fact, it is a stronger relationship. This is an ecological fallacy. If you said this was the student-level relationship, then it would be false, but this says that there is a stronger relationship at the school.

If you look at two students, one is at the school where 100 percent of the kids are free lunch eligible and another is at a school where zero percent are eligible, and

look at their scores, they are probably going to be 1.1, 1.2 standard deviations different. So there was an overall school effect of poverty.

Since this was matched state assessment data set, I ran the same analysis using NAEP and found essentially the same results. Compared to the state assessments in these four states, NAEP was somewhat more sensitive to poverty, but there is the same finding of the school level relationship being 50 percent stronger, or the B weight being 50 percent greater from the individual student relationship.

On to some new data on the student versus schools. I would like to say that I think there are many research questions, as someone pointed out. Federal funding for programs is implemented at the school level or above, so there are many research questions for which looking at school-level data is reasonable.

I wanted to mention, when you talk about school-level data, we have, at least starting around 2002, the subpopulation breakdowns. What we don't have is the individual level scores. I should also mention that not having to require the same data to be reported from each state, the database does include things like scale scores and percentile ranks and median score and raw score means and so on, as well as percent achieving cut points.

I asked Victor Bandeira de Mello, who is continuing to collect the data, what the status was for '04-05, and he said, "Whatever you have on the websites, we will pick it up." We may have scale scores in one state but not in another. The whole concept of this version of the database is to provide information to the public that the public has provided.

Different tests in different states. We have heard a lot about how that causes difficulties in doing analysis. I would ask the question, does it matter. I think this is another missing data problem, an imputation opportunity. This is the way I have viewed it.

Suppose that you have fourth grade scores in some states and no fourth grade scores in another state, but we still want to look at the relationship of some input factors to achievement. If we supposed an extreme, that we were really looking at fourth grade scores to represent schools, then we can take the third grade score as the basis for imputing fourth grade scores. By the way, we are standardizing it; I just use the third grade score.

In fact, as a digression, we found in some of our student level analyses a few years ago that third grade scores from state assessments often correlate -- well, can correlate more highly in one state with the NAEP scores in fourth grade, then in some other state, the fourth grade

scores correlate with fourth grade scores.

When we are thinking about things at school-wide reforms or anything where you want an indicator of the achievement over all of a school, third and fourth grade scores might reasonably be taken as two observed measures of the same latent trait. Obviously there is error involved.

The question that I tried to put together some data on is, how much of an error; does it matter. The question isn't whether percentile or scale scores are the same, of course they aren't the same, but whether the results of the analyses would be the same if a different measure were used.

I think of as a model the t-test, and whether the t-test is significant. If we have an imputed value, another score, that is correlated .9 with the score that we wanted, then the student's t-value is probably going to be about .9 as large in the same range, to derive the formulas. So if we are looking for effects, where you think you can right around 2.1 or 2.0, then .9 is going to turn some of those into non-statistically significant effects. On the other hand, if you have a strong effect for the size of sampling you run, you might have a T of three, six, nine, 12, depending on the nature of the thing you are testing, or where the T is .1, .3, .5, it doesn't

really matter that we have something that is a .9 or even a .7 imputer of the actual value.

So the critical statistic I looked at was the correlation between the measures. What I did a few weeks ago was extract a whole bunch, some 27,000 correlations among the school level state assessment scores. These were pairs of scores which only differed on one attribute. It would be the same grade, the same subject, same year, same school, and two different tests or two different measures, like the difference between a percent achieving a mid-level criterion versus a percentile.

We had in this analysis that I put together some 8,000 correlations that were between different statistical summary measures for the same test in the same grade in the same subject in the same year, and state. On the other hand, only 95 correlations in five states that were between different tests, that were in the same grade, subject, year and state, and had the same summary measure.

We didn't go out after -- and states don't often put on their public websites two different tests for a school. That is why we have so few of these pairs of correlations to look at.

First, different measures. The average correlation between -- I am lumping here; the paper has the numbers for each pair, but I should mention, those are

unweighted least square means, and I am going to revise the paper and make the weighted least square means. This is a more accurate measure I have here as a summary.

The average correlation between raw scores, scale scores, percentile ranks, median scores, normal curve equivalents, accountability indexes which are counting -- like, if there are three standards, one for the first standard, two for the second standard and three for the third standard -- and mid-level standard, it is .95. The standards are more down around .92 or .93, whereas the others are maybe up around .95, .97, .98.

It is critical, though, that the standards be mid-level standards. When I take the cases that have one of the measures being the percent achieving an extreme standard, like the standard that fewer than 20 percent of the kids in the state pass or more than 80 percent of the kids in the state pass, that is not generally as reliable and not as well correlated with the other measures.

So I think it is important that if you are going to do analyses that use these cut scores, I would claim that this .95 is an answer to those who say that we have to have scale scores. It would be great to have scale scores, but if you don't, you have a pretty good indicator in the percent meeting some standard that is around the median for the population in that state.

What about different tests? As I said, we had a smaller database for this. The average correlation between scores on two tests that are the same subject and grade and year, given to the same kids, the correlation, these .95 percent correlations, average .92, I would say quite acceptable for using it as an imputation.

I should point out, I had a sixth state, but when I looked at it, the correlation was .3 or something like that, but it was because it was two different tests. One was a Spanish version and an English version in Texas, and I assume that those were given to different kids. So they didn't enter into this. So in talking about tests, I think they were given to the same kids.

Grades. Now we get into the more substantive stuff. The average correlation between third and fourth grade scores in the same subject in the same year in the same school is, there are some 680 correlations that are between fourth grade and fifth grade -- is .76.

Now a T of two becomes a T of around 1.5, so you may be losing some significance or finding some things significant that aren't. You need to know what these kinds of errors are in order to decide whether you want to spend money or not spend money on the basis of the finding of the evaluation that this was a relationship, whatever it was, to achievement.

But again, if you came out with a T of .1 or .2 or .7 or .8, this .76 isn't going to affect your result. Likewise if it is a three or above, it is not going to affect the result.

I guess what I am trying to do is to provide the background information so that not everyone has to go out and do this. This is based on -- there are 25 states involved in this, not all 50, but it is a whole bunch of data.

The surprising thing about this, the standard deviation of this distribution of correlations is like .05. You can bet money on the correlation. The next one you do like this just differs on which grade is being -- between .66 and .86, two standard deviations.

So anyway, .76 isn't a nonsense number to forget about, because there is going to be some other number in some other state. These correlations -- and many of the numbers I am showing you are surprisingly stable across states.

But why isn't this higher than .76? These aren't tests given to the same kids. If we look at two adjacent years in two adjacent grades, so that we have the same cohort of kids. You might say, wait a minute, some kids might leave the school and some kids come in, so we don't have the same kids. I find I don't agree with that

argument. That cohort is an animal, is an entity, is a unit that progresses. Just as for individual students -- well, first of all you don't usually have attrition of a whole grade in a school, but also, even if you don't have attrition, different kids have different summer experiences, things that add error into the measurement. The fact that there is some mobility will cause some error. For some schools you would have a lot of mobility and find much lower correlations.

Anyway, when I take correlations for a different year, not the same year, but they are for the same cohort, the correlations tend to be more like .8, .81 as the average, with a standard deviation on that comparison of .04. Very stable kind of result. There is a cohort effect that you can see in these.

This is almost blasphemous. Different subjects? Oh, no. Reading is reading and math is math. Well, yes, that is certainly true, but I think we all know that if your kids are reading better they are going to have a better chance to learn math. When we have two different subjects in the same year -- now, these would be given to the kids, we assume, in most cases.

I should say, I have been focusing on grades three, four and five for this particular presentation. We have data that go into the middle school and high school,

but all of these were taken just for grades three, four, and five, and combined elementary, which I haven't talked about the scores.

Anyway, the correlation is what I think is quite high, .86. So if I have reading scores in 40 states and math scores in five states, maybe I can use all 45 if I need to, because my math scores aren't going to be that different from the reading scores in the same school.

Next issue, multi-state analyses, how do we do them. I just want to talk about two of the kinds of analyses I've done. First of all, it is obvious that we want to do multi-state analysis. The NLSLSASD [National Longitudinal School-level State Assessment Score Database, also known as SSASD or SLAD], national because it is all the states, longitudinal because we have the same schools year after year, unlike NAEP, which chooses a different sample each year, school level state assessment database.

We have to start off with within-state analyses because we have a different test. But can we then combine it across? This is something like what Bob Linn was showing you earlier, just a more recent computation of it, how state standards for proficient vary widely. If the federal government says that they are getting consistency by having each state report the percent proficient, that is not true.

Here I have one graphic I have put together that suggests some states have very high standards and some have low standards. On the vertical axis I have shown the percent of the nation's students who pass a state's particular primary standard. Most of the states at this point have multiple standards. In the project that we have been doing, we identified what was the primary standard in terms of what was reported as we understood it, and sometimes it is hard to figure this out, for AYP and No Child Left Behind.

For instance, in Louisiana, they had a standard such that 22 percent of the kids in the country could pass it, and in Mississippi they had a standard such that 87 percent of the kids in the country could pass it. Grade four reading standards, these are.

The way we did it is a little different from the way Henry Braun did it, so we could get a measure of the real standard error of these, which depends on whether NAEP is correlated with a state assessment. To do this we had to go from one state assessment to NAEP to the other state assessment. What I graphed on the x-axis is how much error is introduced because we weren't going from NAEP to NAEP to NAEP, looking at the NAEP. We can do the NAEP to NAEP by using the different plausible values.

Down in the lower right there are six states in

which I concluded that, we will put these in our report, but we think these are estimates of where their standards probably have a lot of error. West Virginia, we were using a composite math and reading, whereas in the other states we were using reading. Nebraska had a bunch of different tests in different schools. Texas set its pass rate, which is what we had available to us, extremely low. So if you look at the scatter plot, they are all up around 100 percent, above 100 percent, and there isn't any reliable variation there. I don't have anything to say about the other three. It is an arbitrary score, where we set the cutoff on the x-axis.

What we have done is to do evaluations in terms of the effect sizes of differences between target schools and other schools in each state. For this we used -- the effect size is a statistic which is the difference divided by the standard deviation of that statistic.

This is a comparison that we had turned into the Department of Education back in 2002 based on 2001 scores. One year, school-wide Title 1, reading gain effect sizes, compared to other schools in the same state. Then we averaged over all of the states. We had 41 states where we did this at the elementary level, 39 at the middle school level and 27 at the high school level, and the effect size was pretty small.

We had five out of the 41 that were significantly positive and one was negative. There doesn't seem to be much of an effect here. I used the word effect. I found it unfortunate that we used the term effect size to talk about the mean difference divided by the standard deviation, because people think you are talking about cause and effect, and some people will extrapolate from cause and effect. All we are looking at is a correlation.

One way to do the multi-state analysis is by doing an analysis in each state and counting them up, or averaging the effect sizes. Another way, multi-state analysis of state assessment scores, they pull within state analysis, they have substantial power.

Years ago we merged the state assessment data with the schools and staffing survey in 20 states. That is all we had at that point for 1994. We did some structural equation modeling to see what correlational relation -- again, this is cross-sectional data, SAS '94 and state assessments in '94.

I am going to show you two partial tables of results. We set the means to zero in each state. Just focus on one row here, looking at the relationship between class size -- it is a path coefficient, doesn't suggest causality -- between class size and achievement measures. You see that there are in two of the levels insignificant

effects, maybe minorly significant at the high school level.

However, when we add in NAEP, it is obvious, you take the state assessment scores for within-state variation and the NAEP state means based on those schools for the between state variation. We can get a different set of coefficients from the synthesized measure. Looking again at the class size row, association between these factors and achievement, and now we find we have much larger path coefficients. What this says to me is that there is more of a between-state effect. If I were going to design something to look at, whether there is a class size effect, I would try to be using a sample from multiple states. There appears to be more of a multiple state correlation with achievement.

In the paper I did for this back in '99, I also turned the thing around and used achievement to predict school climate. As you know if you have run these SEMs, you are just looking at the same covariance matrix essentially.

So I have to emphasize that when I say this, I am not trying to say smaller class sizes cause higher achievement. Here is a correlation that requires some causal interpretation of some type.

Next I want to talk about demographic

adjustments. I am going to run through this quickly. I tend to do the same things that Gary Miron was talking about yesterday. You can find all of these if you go to [www.schooldata.org](http://www schooldata.org) and look for the sub-pages reports.

This is an appendix of a report we turned in in 2002 to the Department of Education. This is one state, Virginia, reading achievement in grade four in 2001, 1,095 schools. What I have graphed on the y-axis is what we had for them that year, which is the school's percentile rank for reading, mean or median, whichever it was, and the percent on the x-axis eligible for free and reduced price lunch.

We were interested in the relationship between any of this stuff and Title 1. So we have the Title 1 schools indicated by red diamonds, and the non-Title 1 schools are indicated by blue diamonds. What we see in this, the money was going to the poor schools. I should say, we have a couple more things in this graph courtesy of Microsoft Excel. We have two straight lines that will draw on any scatter plot that you ask it to.

In the upper right-hand corner, we had what was quite a lot of interest to people at that time anyway, and I think it still is. It is identifying schools that are high flying -- that is an education trust term -- or beating the odds, the U.S. Department's term, schools that

are in the top quarter in poverty and in the top half in achievement in the state. The idea is that one might want to go to those schools and find out what they are doing.

If you want to try to use some user-friendly use of these data, you might go to the Edtrust website and go to Dispelling the Myth Online, where they have used our data set. We produced that sub-webpage, and you can find schools that are beating the odds or high flying.

I am going to skip over some stuff here. You can read it in the paper.

Trend analyses. I have two analytical gripes about using ANCOVA for the pretest because it has error. We should really be using GAIN scores. When you have three or more time points, not linear, please, most educational interventions aren't linear or aren't gradual. If you have multiple time points, come up with the specific hypotheses or research questions that would mean that it is important to look at different categories. If you want to worry about the fact that some of these planned comparisons are not orthogonal to others, but still, you want to look at the ones that are of some importance for decisions you might make, based on these results.

Let me just mention this. For enhancing the database, one of the things that is really of great concern to me is, people are getting in real trouble over

accommodations, putting in accommodated scores or leaving out the accommodated scores. People worked for a long time in the first part of the century getting to standardized administrations of tests, and we need to do some serious research and accurately record the accommodations given, and ed test research data.

Summary. I realize there are technical limitations on this database. Some that are specific to this database are that a school is a unit, and there are different tests in each year. But I think my point of view is that these can be dealt with. We can't answer certain research questions, but we can do other things with the data, answer other research questions with the data. I think the lack of randomization in much of educational research is a much more serious deterrent to making causal inferences.

So what am I asking you to do with these data, or offering to let you do with them? Use it for exploration. We are finding schools with particular attributes that are achieving well, for exploring what are the relations to expect on things. That is my second one.

When you go out to plan a study, you don't have to guess at what might be the background statistics on achievement. Here we have a database on achievement at the school level involving all of the states. So we can

generate expectations for outcomes and for relations of treatment, two outcomes; look for those natural experiments that are going on using the data on these 80,000 plus schools.

Finally, I would say we can use it for demonstrating correlations. We can't demonstrate causation. We can demonstrate correlations that by golly, call for the causal explanations. I think the RFPs for evaluations and research should be much more emphasizing that proposals should come in dealing with these 35 or however many it is alternative explanations. If you think about it ahead of time, maybe you can design the study so that you get a bit of an idea whether this or that or the other factor is really what is causing the difference.

With that, the green vegetables and the proteins, here is a little visual dessert, a picture of sunset over the Tufas in Mono Lake in California. Thank you.

DR. YEN: Wow, Don, you really packed a lot into your half hour. Questions, comments?

PARTICIPANT: I've got one about the correlation. In the beginning you were talking about the correlations of measures and so forth. That was over the full range of achievement, right, from the very lowest schools to the very highest schools?

DR. MC LAUGHLIN: Yes.

PARTICIPANT: I would imagine in most program evaluations, they are dealing with one part of the distribution, typically the low end. Is that correlation the appropriate statistic to be looking at to see whether these measures are comparable, or should we be looking at a correlation that is restricted in range?

DR. MC LAUGHLIN: It depends on what you are asking. If you want, you can ask is this a good substitute for that. You might look at the values of the whole range, but the correlation on a restricted range is going to be lower because of the restricted range.

On the other hand, many times people are looking at regression weights rather than correlations, and the regression weights are not affected by that restriction of range.

PARTICIPANT: The issue is a little more complex than what you thought it might be. Some of the schools will be testing the same children in English and Spanish if they have dual language programs, and so they are interested in language outcomes in both languages. Others will be testing different kids in the dual languages, depending on the nature of their bilingual equivalent.

DR. MC LAUGHLIN: Well, that certainly wouldn't affect the correlation.

PARTICIPANT: Don, you said that you did the

analyses that you presented early on on grades three, four, and five, where you found those really high correlations. What would you hypothesize to find if you looked at the higher grades?

DR. MC LAUGHLIN: You have to be careful that you are looking at the same schools. In some states, between five and six is where most of the schools switch from elementary. I would need to very carefully look at that.

The other set I would be willing to look at quickly is seventh grade versus eighth grade, because seventh and eighth in most states are in the same school. Again, if we have tenth, 11th and 12th -- although I think the quality of the tenth, 11th and 12th grade scores on the database are lower than the rest, because what many states are using is the high school exit exam, or some exam that kids can take multiple times. So we have not made a great deal out of the high school data in this.

PARTICIPANT: (Comments not caught by microphone.)

DR. MC LAUGHLIN: I ran this a couple of weeks ago for this and didn't find any results that were different. You are right, I presented some results showing how the school level and the student level results, how they are similar but some are different.

I have a feeling about some of these things, that

one can do the same analysis to get at a research question that people understand better sometimes, just doing straightforward regression at the particular level you are interested in. I understand that if you are looking for the range of beta weights in different states, sure, you run a version of HLM. But no, I don't have anything special to tell you about that.

PARTICIPANT: (Beginning comment not caught by microphone) -- graph that showed the placement of the state standards, and then some measurement of error. I wasn't quite sure what you were calling error. Was it based just on local relations with NAEP and the state scores, or were there other things you were looking at that led to your estimate of error?

DR. MC LAUGHLIN: Other things I was looking at. What I did was to use the state-level linking that I had done to reproduce from NAEP what I think the state would have reported for each school. I had what the state reported for the percent achieving the level, and then I came up with what level matched on NAEP, and then I look at the NAEP data. Overall I had that, but for this school I might be off. It won't reproduce that percent.

So I looked at how much error there was in reproducing, and then I compared that amount of error to the amount of error I would expect if it was just NAEP to

NAEP. I am basing this on a sample in each school, and we have standard errors of measure.

So the one point, the one way at the left-hand side, would be if it was NAEP versus NAEP linking. So it is farther to the right if there is something different about the test compared to NAEP.

DR. YEN: Thank you. Last, last, last.

PARTICIPANT: Don, would you be prepared to say that that was a quantitative index of alignment?

DR. MC LAUGHLIN: No. I think a lot of people are talking about comparing paragraphs and words that are descriptions of curricula and descriptions of standards, and that is different. I am looking at the statistics.

DR. YEN: Thank you.

**Agenda Item: Session 7: Synthesis Discussion**

**Panel: Balancing the Policy and Technical Issues**

DR. DUNBAR: We are doing great on time, aren't we? Very good, everybody.

Time now for our synthesis panel for the end of the day. We have four speakers who have been discussants, who will become speakers by the time we are finished, who have been listening attentively to today's presentations. They are going to each make individual comments. We have about one hour scheduled for this session, and then -- oh, that is a fine revision, five minutes for my closing

comments.

What I expect will happen is, we will hear from each of the discussants, and then we will expand the discussion to the group. I will try to come back with a few more general consensus type comments that I think reflect the issues of today.

So we have David Francis from the University of Houston, Diana Pullin from Boston College, Bill Schafer from the University of Maryland, and Laurie Wise from HumRRO.

DR. FRANCIS: Thank you, Steve. I'd like to thank NRC and the planning committee for the opportunity to be here. I really enjoyed the papers, and enjoyed the presentations. The papers, if you haven't had a chance to read them, are really outstanding and very informative. I'm not going to comment on them individually, but I recommend them to you. I thought they were very informative.

It is clear that the SSASD is a great resource for educational researchers and program evaluators, and it is a great step forward in terms of developing what I would call an analog to the public health data system, but for education. Clearly it is not complete, and I don't think any of us would argue that it is. I think that Gage [Kingsbury], in his presentation, highlighted some of the

key strengths of this database as a resource that we didn't have.

In terms of the workshop, Stephanie [Stullich] and David [Goodwin] asked us several questions: How can we use the SSASD to evaluate federal programs?. Some approaches are better than others, what kinds of inferences can we support? I am going to try to focus on those issues in terms of what has been talked about.

I will just say at the outset, I don't think I have anything unique to say. Everything I am going to say has been said already, so what I am really doing is trying to highlight those points that I have heard reiterated over and over again that I think make some sense.

It seems clear that the answer to the first question is that there are many ways in which the SSASD could be used to evaluate federal programs, but precisely how is going to be necessitated by the program, or another way to put that, the question that we are trying to answer.

I think the idea that there is a single best way in which to model the data that exists within the database is a fallacy. I don't think that that is true, and I don't think we should search for that holy grail. The strongest approach is going to depend a lot on what the particular question is, and how to fashion information.

So I want to repeat a little bit of what Judith

Singer said. She made several excellent points in her top ten list. I think her point number zero, which she didn't number as point zero, is that the best analysis is going to start with a carefully articulated question. It is really clear that when we think about program evaluation, we have to start out by defining very precisely what is the question that we are trying to answer.

Then the second step that she articulated, her point one, was a carefully and fully articulated statistical model. I would insert somewhere in between those two, that is, between the question and the statistical model, we need to think as well about the mechanism of action for the program, how is it that the program is going to impact on student learning. We have to think about how we are going to incorporate that into the data that we are going to try to have access to, and into the statistical model, how is it that that mechanism of action takes place. If we don't have a good strong idea about what it is that the program is impacting at the instructional level or at the student level or at the school level and how that then translates into achievement, our models are going to be fairly vacant, and our ability to get at causal inference is going to be that much weaker. We are already dealing with generally a weak approach because we are dealing with observational data.

I think in terms of articulating those mechanisms, we need to think both about how the program would have intended consequences in terms of the intended mechanism of action, but then also unintended mechanisms of action. We heard some about different kinds of things, like people gaming the system, cheating on test scores and things like that, which will also have an impact on achievement that we need to try to disentangle the extent to which those things are operating.

It seems that in all likelihood, when we go to do a program evaluation that incorporates the SSASD, we are going to need to pull in additional data to assess how these mechanisms of action might be operating. Certainly one thing that Michael Scriven talked about was going out and sampling schools for more intensive data collection. I think that is clearly a strong option.

The other side of that coin, though, is that there are a lot of local evaluations that are taking place of programs. Those local evaluations, where the designs might be somewhat tighter in terms of the data collection, need to also access the SSASD to try to put the achievement outcomes that they are finding into a broader context of how the state might be operating. So I think that sort of sampling can go in both directions, and utilizing the SSASD to improve those local evaluations I think would be

powerful.

I think that from the standpoint of -- and this has been touched on by several people -- one of the strengths of the SSASD is this ability to draw on longitudinal data collected within schools at the state level long before the program gets started. So having multiple years of pre-program data really improves the strength of the designs and models that we can fit with respect to trying to understand the causal impact of programs. We have to exploit that.

I was talking earlier with Elizabeth [Stuart] about ways of going about trying to establish what the post-program expectations might be in the absence of program effects, based on all of the prior year data that you have. There are many ways that you can do that, and what you should be doing is thinking of multiple ways to do that, so that you are not dealing with one set of expectations, but potentially multiple sets of expectations against which to compare actual performance.

So I think when we think about statistical models and approaches, Judy [Singer] laid out a nice set of rules for people to follow in terms of articulating the model, consideration of omitted levels, omitted variables, looking for alternative explanations, conducting sensitivity analysis. You can encapsulate all of that into following

the best statistical practice for program evaluations.

So when the Department [of Education] is taking into account what different contractors are telling them about how to utilize the database, these are the things that they should be looking for in terms of statistical practice, and not did they use the right model, necessarily.

I want to say a couple of points about how else the SSASD might be used. I mentioned the issue of using it to inform local evaluations. Certainly, from the standpoint of design it can be used to improve our subsequent designs. The Department posed the two groups, the PPSS, which is doing evaluation work, and IES, which is favoring randomized designs. But it seems to me that these two organizations should and could work together, and that the utilization of information in the SSASD for designing randomized studies where the number of schools is going to get randomized is fairly small, and is a real strength. If I can match schools up based on five years of pre treatment data and then randomize those schools after matching to treatment and control in a smaller program, I can get a much stronger design.

So what can we do to strengthen the SSASD? That was one of the other questions that we were asked. A number of possibilities have been discussed. One is

student level data, one is program information, and then the other is scale scores. I won't say anything more about scale scores; I think they have been talked about a fair amount, and the importance of incorporating them.

I do want to say a word about program information. We touched on this today, a number of the speakers did, but the whole issue of standardization, not standardization of the achievement data, but standardization of the data collection practice.

There is a fair amount of consistency within states, but as you start moving into talking about collecting program information data, like what is going on in the classroom, what is the curriculum that is actually being followed in this particular school, there is not going to be a lot of standardization about that. If we are going to make use of that data in terms of analyzing achievement data, I have a lot of concerns that if we don't pay close attention to how that data gets collected, those analyses won't be very meaningful.

I do want to talk about the student level data issue. I think it is clear that there are benefits to it. I think Don's points are also worth taking into account. Where I get concerned is the idea that we might try to actually pull in that information into the SSASD and create some monolithic mega data structure, which I think would be

a tremendous cost for not too much utility.

I think the ideas that were talked about earlier, about linking into the existing longitudinal database, is a much stronger idea.

I'll stop there.

DR. PULLIN: David and I did not plan ahead together, but I am going to pick up on his theme and ask you to focus on what questions you are trying to answer with the use of this database or any other.

I think we have had a really provocative and informative set of discussions over the past two days among some great thinkers from the scientific community. I am not a scientist. I spend a lot of time, though, thinking about how science interrelates with public policy concerns and legal requirements, and I want to talk about these issues from those perspectives.

I thought in particular I would like to make sure that we don't lose track of the point that Judith Singer made so beautifully yesterday, that the real issues here are how our discussions inform efforts to improve enhanced opportunities to learn and real enhanced educational achievement for all students. I have spent enough time hanging around with really good psychometricians and statisticians to know that they are excellent at figuring out new algorithms to constantly try to propose different

ways to answer questions and solve problems, but that it is sometimes easy to lose sight of the larger issues.

So I thought it would be useful to bring you back to what some of the people from the Department were talking about at the very beginning of our meeting yesterday. They talked about this GPRA animal, the Government Performance and Results Act, which was put in place in 1993 to try to impose accountability standards on all the federal agencies.

I think it is important to understand that in that law, the Congress articulated definitions of what are outcome measures and definitions of performance goals and definitions of what constituted program evaluations that would meet these new legal mandates that were being imposed on federal entities. Among these requirements were very clear requirements within the mind of Congress that federal agencies should be held accountable for setting targets for themselves, for articulating levels of performance in a tangible, measurable way, against which -- and this is the language of the statute -- against which actual achievement can be compared.

So when the No Child Left Behind Act came along and Congress contemplated how to write that statute, and the Department of Education contemplated how to implement it, clearly part of the goal was to try to create a

defensible mechanism for program evaluation for such things as the very large amount of federal funding going into the No Child Left Behind Act.

While we have had a lot of discussion in the past day or so about this big press toward randomized trials and the like, I want to remind you that Congress also in writing the No Child Left Behind Act wrote a definition of scientifically based research that is broader than many people remember in some of the conversations that we have. There is ample room in there for other empirical approaches, as long as they meet some criteria of their own about how systematic they are, and there is in additional room for observational approaches that don't include just randomized trials.

So as we struggle through thinking about how to address the questions that have been raised about these databases, I think it is important to ground ourselves in the notion that part of what has been very unclear in our discussions of the past day and a half, as we have talked about program evaluation is what really is it that we seek to evaluate through the use of this database or any other.

You can come at that question from a lot of different perspectives. You can raise the question and think about answering it not only in terms of the scientific data that have been talked about in our meeting,

but also according to some other criteria that are set out in NCLB elsewhere. Among other things, the Congress asked that people pay attention to relevant professional and technical standards. Maybe it is just because I spent so many months and years re-writing the last version of the program evaluation standards and then foolishly agreeing to rewrite the testing standards, but part of what I think we need to make sure we are paying attention to are those relevant professional and technical standards, in addition to the kinds of scientific considerations, or coupled with the kinds of scientific considerations, that people have been talking about here today.

So when I think about how to use the database that has been the primary object of our discussion, I wonder when we talk about program evaluation exactly what kind of program we are talking about evaluating. Just within the No Child Left Behind Act, there are many different possibilities of programs that need to be evaluated. There is Congress' primary interest in proficiency and the movement toward proficiency for all children in 2014, but there are also important program evaluation questions that could be asked about the reading interventions that are also funded through the statute and its Reading First initiative, about such AYP factors as dropout rates, which have come up here in passing, about

such issues also addressed in the statute as the pursuit of improved teacher quality, or the alleged pursuit of teacher quality, depending on how you react to the question.

These questions are scientifically difficult, no matter which of them you choose to pursue. Certainly some of those scientific difficulties have been highlighted here today and yesterday, as have some of the possibilities. But there is also the need to recognize that the United States Department of Education in its implementation of this statute has engaged in the exercise of a very broad set of authority it is given under the statute to do a number of different things, all of which might relate to any of the program evaluation questions you might choose.

In some instances, the Department has exercised its broad authority in very clever and extremely effective ways with a speed I have never seen before in the years I spent either working in Washington or studying what goes on here.

Some of what the Department can do, for example, is change very quickly what local building administrators think are acceptable approaches to teaching reading. I have a lot of school principals in the classes that I teach, and I can see them come in and work to very quickly change the kind of reading programs they have in their buildings, in direct response to key decisions made at the

Department of Education here, some of them not necessarily ones that the reading experts would agree with, but decisions that nonetheless had a rapid impact on practice.

Similarly, I can see the choice of an item on a state assessment test having consequences within 24 hours in what my principals inform their teachers they should do in something like an elementary curriculum. So there is a lot of power within the Department and within the states to address many of the issues that we have been grappling with today.

Part of what I would encourage further thought on is not only these scientific questions, but also the questions about how to coordinate within the federal agency and in the states and in those two levels of government working together, approaches to issues that could in fact inform the acquisition of additional data that would strengthen the database under discussion here today.

So for example, the Department's power to approve plans under Title 1 of the No Child Left Behind Act, which is the largest source of funding that goes from the federal government to the states and localities, includes a considerable power to declare what is acceptable as state and local practice for the receipt of those funds. The peer reviews that are used to evaluate those state plans or workbooks have been mentioned here several times. That

peer review manual, that I suspect Bill [Schafer] will talk about in a couple of minutes, is an important opportunity for the government to provide assistance to those who want to add to the possibilities for further enhancing these databases and their usage.

Similarly, I assume that these new state longitudinal grants that were described earlier today also could provide an important opportunity for the federal government, if it wished to do so, to impact in a more significant way state and local practice.

Now we come to the part that is both a legal problem and a political problem. There is a real federalism issue here, a real strong set of potential problems that have already come to light in many different ways and will become more severe over time, over the relative power of the federal government compared to the states and the local education agencies.

One of the quiet things that William Rehnquist did before he passed away was to begin to completely redefine from a legal perspective the relationships between the federal government, the Congress, the courts and the states. There will be more of those sorts of disputes in the years to come.

But this also becomes a political problem for the agency here and for the various states, because the No

Child Left Behind Act has to be reauthorized by Congress, and that reauthorization is currently expected to happen in 2006 or 2007. That is about the point in time when there start to be even more data available at the state level about progress under either the AYP standards or the implementation of the high stakes sanctions on kids, teachers or school buildings.

That suggests that there is not a lot of time between now and when Congress considers this statute the next time to engage in the kinds of changes that have been considered in the past day and a half of discussion here. But it certainly suggests that this is an opportune moment for researchers to try to frame the sorts of political responses or technical suggestions to decision makers of the sort that Judy was talking about yesterday.

So I would in the brief moment of time left to me say that I think there are a lot of possibilities here to make good use of these databases, to enhance these databases, and in a thoughtful way to acquire more useful and meaningful and perhaps not highly intrusive collections of data to inform whatever program evaluations we are going to undertake.

For people who have asked me questions about such potentially thorny issues as whether gathering student data will present a real hornet's nest of difficulty from a

legal or an ethical perspective, I would say I am usually one who sees legal trouble, but I would say that in this instance, based upon some past experience I have with the Department and with NCES in the 20th century, there have been instances in which the Department has gathered databases like NELS, run into difficulty later, particularly because of the Federal Privacy Act that limits intrusion or limits data collection on the part of the federal government, where we were able to encourage the construction of some not too intrusive and easier to implement, if you start doing it from the front end rather than in the middle, kinds of ethical and legal protections that could ensure that you could collect these data, you could protect student privacy and identity, you could be able to get information about these critically important small subgroups of children who we have had to remove from reporting because they could be identifiable to the public, but give ourselves an opportunity to know even more about what is going on in the implementation of both the federal and the analogous state statutes.

DR. SCHAFER: Thank you. I promise not to talk about the peer review process. I have said enough about that.

When I gave my very first AERA presentation, it was at the last session. I was the last speaker. People

came into the room -- it was crowded, but they came into the room with their suitcases, and by the time I spoke there was hardly anybody there. I would like to particularly thank Laurie Wise for being Eva [Baker]'s substitute to save me from that fate.

It is certainly a struggle to try to find new things to say after all that has gone on here. So maybe I will repeat some of what has been said, but I'll try to add a new twist here or there, and maybe one or two things that I talk about will be new.

A useful exercise in trying to improve the database would be to determine who might have a need for it, what sorts of questions they might want to address, and what their data needs would be for those questions. One anticipated use of the database is to evaluate educational interventions. We have heard a lot about that here.

As it is currently designed, the database seems most helpful to persons who know that certain schools are implementing certain programs, such as a district which is evaluating certain of its own schools, or an externally funded program that has identified its implementation sites.

But say a school's personnel needs to survey the results from interventions in schools that are like theirs demographically. They can see if and when other schools

improved, but they have no way to investigate the programs that may have led to more or less success. Or there may be interest in the success or failure of programs with various combinations of elements that exist in more popular interventions. This suggests it would be helpful to find a way to describe the educational programs in the schools, and especially to describe what was done differently year by year.

Perhaps some sort of written description could be contained in an external file that can be easily linked to the present database. It is a mammoth undertaking. Other ways might be easier to accomplish, but even if this could be done for a sampler of schools, it should make the database more helpful to at least one class of user.

Some powerful growth models have been described, but to what extent can they be ascribed to the data in our database? We have heard discussions about using data at the school level or at the student level.

I think we should be thinking here along three lines. The first of these adds a me-too to David Thissen's support of Judith Singer: What assumptions do we need to make about student level data in order to learn what we can from student level growth models when we have only school level data. If we can answer this question, we can evaluate the assumptions with other data in a coordinated

research effort, which is an idea that Donald Rubin suggested earlier today.

The second line is, are there any feasible additions to the database that would enable using student level data. Failing that, are there ways to coordinate and otherwise facilitate access by researchers to public state level databases. The state-level panel had some comments about that.

The third line is whether we can include information that affords inferences about student-level results without student-level data. At least one paper and some comments described lack of data availability on variability in the database. Consistent with the APA [American Psychological Association] publication manual, we should have sufficient statistics for everything that is reported in the database. For example, this means for means, we would also have standard deviations and sample sizes.

Predicting future growth from past growth may be hazardous. The conventional wisdom is that assessment changes result in brief growth on the new measures with less positive results thereafter. We are now seeing a large number of new assessments as a result of NCLB, and can anticipate seeing growth for a few years as a result. However, linear projections of growth may over predict and

make schools look on track when in fact they are not. So perhaps we should be thinking about the sustainability of growth and whether growth targets are realistic. This will be a more important question, especially in states following Ohio's lead that have backloaded their AYP targets.

As we move toward 2014, all states will feel increased pressure to make changes in the criteria for AYP. This may change beliefs about which states are setting appropriate cut points for their proficient achievement level. In the end, this is a policy issue rather than a technical one, because its legitimacy depends on impact rather than understanding or insight.

Indeed, we may eventually conclude that states with the most lenient achievement targets have the most positive influences on their educational programs. It would be interesting to see the correlation between NAEP scale cuts for the proficient achievement level and eventual growth on NAEP as well as state tests.

As has been mentioned in earlier comments, we are not like other countries, in that we do not have a national curriculum. Instead, we ask each state to set its own content as well as proficiency goals. What is the role of testing in this environment?

I have been thinking recently about what I have

called the fundamental accountability mission, which is to test every student on what he or she is supposed to be studying. In fact, the test blueprint can be used to define these curricular goals. If these content goals are indeed different, then why would we ever want to link the state test to NAEP or to each other? I think Bob Linn's discussion raised a lot of questions about the value of performing those linkings. Especially when we consider the effects of a testing program and what goes on in the schools, even NAEP might be a negative influence on what a state education agency is trying to do. Fortunately though, the content differences aren't that severe, and some studies have shown this. As a matter of fact, the NAEP frameworks are often used as justification for movement in state content standards toward those frameworks.

A problem researchers should be aware of is the existence of arbitrary maxima and minima in student data. As I'm sure everyone here knows, item response vectors with all zeroes or maximum scores tend to indeterminate theta estimates. Therefore, states set arbitrarily values for the LOSS, the lowest obtainable scale score, and HOSS, the highest obtainable scale score. HOSS doesn't occur much, but LOSS does, and sometimes frequently.

Since states set these values arbitrarily, they

can vary across states, and even across grades and contents within states. Additionally, state policies differ on whether students who are absent for tests for various reasons are declared missing or are assigned to LOSS.

A problem with LOSS is that as an extreme score, it can disproportionately affect means. Many states have come to use percent above cut instead of mean to report data, but if a researcher wants to compare score data rather than counts of students, a better choice might be a median or some other percentile. These can be indexed and have standard errors, and so can be used in meta-analyses.

What do we learn from one study? Much has been made about randomization to conditions. Failing randomization, we have heard about matching covariates, propensity scores among other ways to introduce controls for selected variables. We have also heard about using chain scores when students are their own controls. But we still only have one study.

Perhaps we need to go beyond the one study paradigm and consider where that leads us. If we expect to base important decisions on scientific understandings, perhaps we should develop criteria for the support the understandings need to have.

Should we require that studies supporting our crucial understandings be replicated? Should we require

that they have survived a peer evaluation process? Should we require that there have been serious attempts to falsify them? As we all know, we act every day on assailable understandings about what are really hypotheses, so our focus should not be on establishing what evidence is unassailable, but on establishing what evidence is sufficient.

How are we to use non-experimental data meaningfully? One thought is to use it for hypothesis generation and then to test hypotheses with independent studies. The concept of propensity scores as a means of inferring causality from non-experimental studies has been proposed. I know many professionals agree, as I heard one researcher that I respect say, that propensity scores is the only way to infer causality, lacking randomization.

I have a couple of problems with that. First, propensity scores will not necessarily control all relevant contents. There is a counter example. A program works because the principals who chose it were excited about it and conveyed their enthusiasm to staff. There is no way a researcher will include principal enthusiasm as a covariate to establish propensity scores.

Another problem is that I think there is a better way to infer causality, and that is the development and testing of theory through replication. Even absent theory,

multiple independent replications even with pre-post data in individual schools and study sites with school characteristics to study variables seems to me to be at least as strong an analysis as opposed to propensity scores. This to me suggests that meta-analysis of effect sizes within schools is an appropriate research paradigm.

Most often, effect size is thought of as between treatment and control conditions, but effect sizes based on change over time within or across cohorts might be a way to consider to make the database more useful. Although the power should be ample either way within cohort change, it will also yield smaller standard errors because the samples would be related. But I am not necessarily suggesting within as opposed to across cohort approach.

Effect size for a one-year gain is obvious. But as we have heard, effect size could be defined in terms of trend components, allowing meta-analysis of growth patterns. That would allow interventions to be compared across states, as each school becomes a study in the meta-analysis, and by using meta-analysis generalizability of findings across states or school level characteristics can be hypothesized and studied.

Effect sizes such as these have been criticized as being indexed to structurally different variabilities. Even using state variabilities as an approach for

adjustments has been questioned, because state variabilities differ. But I don't want to give up on this analytic approach so easily. Perhaps adjusting differences in state variabilities using ratios of state variabilities on comparable NAEP scores, or school variabilities on comparable state scales, could provide a viable approach to the problem, or state variability on NAEP or school variability within a state could be used as a conditioning variable in the meta-analytic model.

Varying operational definitions of treatment and control conditions has been discussed. If unevenness of treatment application or control applications produce differential effects, that will show up as heterogeneity of effect sizes in meta-analysis. When heterogeneity is observed and all explanatory variables have been exhausted, a common follow-up is to look at outliers around the complex model. This may or may not identify unusual treatment or control abnormalities in specific applications.

Finally, what are we going to do about the policy shift away from a scale metric? If anyone wants to start a grass roots effort, I'll join, as would many others at this conference. But failing that, we can try to influence the design of the database so that it facilitates links to state databases with the needed data for our research, as

Gage Kingsbury suggested.

Thank you.

DR. WISE: I, too, apologize if my comments are a bit redundant. It probably means we all were watching the same symposium.

I am going to try and talk just briefly to summarize what I heard about three topics that I think are salient to the theme of this. The first was basically the theme, how can SSASD be used in evaluations of federal programs. The second theme I am going to talk a little bit about are, what are some of the limitations of using this database in these evaluations, and the third is, what might be done to enhance the database to limit the limitations.

I'm not going to talk a lot about the causal modeling, but I do want to say that I thought there were some excellent presentations and discussions about causal modeling that I hope someone will follow up on and try to get maybe a publication or something that talks about the things that Elizabeth [Stuart] started us with. I am especially going to take away what Michael Scriven calls the elimination model as an alternative to just strict statistical testing, although I like to refer to it as the autopsy model, because that is when I woke up and it got my attention, is when he talked about autopsies.

First of all, how can the data be used? Several

people talked about the value of using the data about the universe of schools in this country as a sampling frame for identifying matched samples that you might then go in and apply your treatment and your outcome variables to. That doesn't suffer from the limitations of having to depend on the data itself for everything you want to know about the implementation of the program. It does allow any of your causal modeling approaches that you would like to do in order to be able to attribute effects to the program interventions. It is a very powerful tool.

It is also, as Gage pointed out, powerful because it has got historical data. So there is baseline data on what student achievement was like for maybe several years, leading up to the point at which you are going to do -- or at which the intervention for federal program was introduced.

The rub comes in when we start to bridge into the limitations; to what extent can you use the school level student achievement means as an outcome or the outcome measure for doing evaluations. There, we heard a lot about limitations and concerns, although we also heard, don't let the perfect be the enemy of the good, or, what was actually said was, an approximate answer to a good question is better than a precise answer to a lousy question, paraphrase.

So what are the limitations then that were talked about in terms of using the achievement means as outcome indicators for doing the program evaluations? First, there is a lot of talk about the validity of the test scores in general; do they measure what the program was trying to implement, and do they measure it reliably and accurately, or are there ways that the scores are distorted or less than perfectly reliable.

In talking about the problems in linking together different scores, especially if you are wanting to make cross state comparisons, or NAEP versus state, is population invariance from the linkings. I was actually part of the Uncommon Measures group [NRC report], and we went through a lot of that. It does seem like a key population that you are worried about the variance or invariance of is individual schools, so would individual schools line up the same way if you used one measure versus if you used another.

So I thought there was some nice balance to the workshop here, because there were a lot of potential issues and problems pointed out, but there was also some data that Don [McLaughlin] presented that I think showed the level of approximation that might actually be feasible.

One of the numbers I am going to take away is the .86. The .86 is the correlation at the school level of the

mean score in one subject versus the mean score in another subject. We worry and wring our hands that math in NAEP and math in the state are measured differently, but reading in the state and math in the state are measured differently even more, and yet there still is a pretty good correlation.

I think one of the problems that we as researchers confront in thinking about using school level data is that we are used to using student-level data. So the lack of precision of information when you have a pass-fail versus an actual score is pretty significant. But again, Don's results showed that if you are aggregated at the mean, so that you are not dealing with a pass or a fail, but you are dealing with a percent pass, which is a continuous metric yet again, you get a pretty high correlation between that and what you would have if you had the more complete data like school means.

So I throw that out as, again, there are both limitations and there are some limitations to the limitations.

One of the other limitations besides the validity of the test scores in general is, we can't really use this to evaluate below school-level programs at this time. These are school-level means, and we are looking at interventions at the school level or higher. It is also

not yet possible to assess individual student growth with aggregate data like this. Some of the wonderful models that Yeow Meng [Thum] talked about, we would have to work at to be able to use these data in that context.

One of the other limitations, and then let me get to the enhancements, is the question of timeliness, how soon will these data be available. I would give the example, that NAEP uses the Common Core of Data [CCD] to build the sampling frame each year to select the schools that will be assessed that year.

But because the Common Core Data is two years and sometimes three years old, they also have to have procedures for identifying schools that were in the frame but closed, and schools that have opened up that weren't in the frame, and to make some accommodations to make sure that they have represented today's school population, not the school population of two or three or four years ago.

The data that AIR has put together has been available in a reasonably timely manner. It remains to be seen with EDEN and so on whether they will get bogged down in the perfect being the enemy of the good, and whether they will be able to produce data in a timely fashion.

I just want to come back to one of the other limitations that I had written down, the between state comparison as particularly problematic. If you want to do

comparisons within year and within state, and you have got every kid taking the same test, that is pretty reasonable. There are still validity questions, but as Mitch [Chester] mentioned, it is sometimes important to learn what other states are doing.

A lot of the policies, both the implementation of the policies and sometimes the policies themselves, are really state-level effects, so you need between-state differences in order to begin to tease out state-level effects, and that is a difficulty.

Let me turn to some potential enhancements that I heard people talking about. I think we all said the more data you can get on the characteristics of the test, maybe it is the meta-data that the woman from EDEN [a Department of Education staff member] talked about, the meta-data about the characteristics of the test. Clearly that is going to be useful, and people need to understand and track when an assessment was changed, what were the characteristics of the assessments, what were the cut scores, et cetera.

Gage and others talked about the characteristics of the programs and their implementation that we might want to do. I think that is a good idea up to a point. We won't know in advance all of the programs that we might want to evaluate, and implementation data probably ought to

be collected at a greater level of depth than we are likely to want to provide on all the programs in this overall database, so the idea that you could link in additional data on samples of schools and on particular implementations I think is a good one, and you should think of it has a multi-stage approach, and not try to put so much burden, too many ornaments on the Christmas tree, is what we used to say about NAEP, but it is similar here. If you make the database want to serve too many purposes, you will never get it done.

The individual student level data would be another enhancement. I think the program that Kashka [Kubzdela] talked about, where money is being given to states to develop these databases, let that mature for a few years, and then there is really a potential for the states having something to share that is in a mode that would be useful and somewhat standardized. So I would revisit that. We may not be quite ready to do that now.

Another thing we might think about is to make sure that wherever possible, we have data on other outcomes besides just achievement scores. I forget, one of the speakers earlier today talked about educational attainment, dropouts and other things as desirable outcomes, and that many of the interventions are targeted towards, not just achievement scores. So to the extent that we can get

multiple indicators of important educational outcomes, I think that is extremely useful.

I want to say that the NAEP data that Don [McLaughlin] has been talking about, and the potential of using it to adjust for between-state differences, I think has very great potential. I was a little concerned; there was some comment about the duplication of effort between the NAEP work that is ongoing and what EDEN is doing, and I just wanted to reinforce the value of having different people working on different parts of the problem. So both the data that Don has put together at times to get samples of data on individual students so you could compare individual and state level results, and certainly all the work to link NAEP results school by school, which most people can't do because NAEP doesn't let out the school identifiers easily, with state results, is useful.

As many people have pointed out, linkages will probably change over time. They will change in part because the state testing program changes. So doing that on a regular basis would allow you to monitor when something is different about the state assessment versus NAEP, so long as NAEP doesn't change too much, and that is a discussion for another day.

I would also urge the Department [of Education] to pay some attention to what is in it for the states. To

the extent this is a collaborative effort, and they are seeing value out of this, you are likely to get not only more cooperation, but more effort on their part to assure data quality and so on.

Then finally, consider the idea of distributed data collection, so that you have the main database focus on a core set of data that you can collect well and in a timely manner, but that you also support and encourage a variety of other collections of data about federal programs in various states, about certainly the NAEP relations, and about as many other things as may be useful, that you can then link in on an as-needed basis.

So with that, I think my time is done. Thank you.

DR. DUNBAR: Thank you all. We have a little bit of time left. I have a couple of cleanup activities to take care of. But first of all, are there any comments or questions for any of our panel here?

PARTICIPANT: One issue that has been raised by some people, but because they raised it, it caused me to reflect on it just a little bit.

There hasn't been a whole lot of attention to the independent variables here. I don't know whether this is something that everybody is assuming that we are going to be able to get what we want. We talked a little bit about

more information on implementation and what have you, and maybe this is not a comment for this panel, maybe it is a comment for the NRC committee to possibly just confirm what people seem to be assuming, namely, that we will be able to get okay data that will be useful to explain the achievement data that we have now gone through in some detail.

Clearly we are going to have data from EDEN. Is that going to be enough to look at the implications of interventions that are brought about because of NCLB? Obviously there are lots of other databases that could be joined. It might be useful to look in detail at some things that might be on the horizon, and possibly interesting analyses.

PARTICIPANT: Let me just make a brief comment on that. We are really in a situation where we commonly use qualitative inference as opposed to quantitative inference when we ask what the relevant independent variables are. We don't know what they are, so we don't know what to go out there and measure. We might try to think about ways in which we could develop a qualitative database.

PARTICIPANT: Let me just say that verging on what was in the Common Core Data is a useful first start, because a lot of thought went into what should be collected and what should be kept in the Common Core Data about the

school characteristics that might be some of the important independent variables.

PARTICIPANT: (Beginning of comment not caught by microphone) -- used for another stratification variable. That is one thing that has led to our trying to hurry quickly to have the data ready in a timely fashion.

PARTICIPANT: I'd like to make one comment about the policy stuff and the role of randomized experiments. One of the rules of randomized experiments is not just to conduct randomized experiments, but also to serve as a template for how observational studies should be analyzed. So if you are faced with an observational data set and you are trying to assess the causal effect of an intervention, you should structure it so that you think about how should parallel randomized experiments be done.

The parallel randomized experiment would have to have very distinct phases. I think Liz [Stuart] mentioned this very briefly. There is a phase of design of randomized experiment which is without any outcomes in sight. We don't get to do randomized experiments, look at the answers, ah, let me re-randomize, see if I can get a better answer, more favorable. Let me do a block on sex this time, I'll do another randomization. You don't get to do that. You get to do one randomization, one design without seeing the outcomes, and then you have to live with

what happens.

How many analyses of observational studies closely follow that parallel? Is there a distinct design phase where you have all the outcome variables locked in a closet and you can't see them. You try to structure the data so that you can do cause linkages.

That is the powerful idea about matching propensity scores. They don't involve the outcomes at all. So you get to design the observation study in parallel with the template randomized experiment that you have in mind, commit to the treatment control groups we balance, and then live with the answer. If you can't get treatment control groups that are balanced with respect to what you have, give up. You can't make a causal inference without making heroic assumptions.

But most of what I have heard today bundles those two phases together. They haven't distinguished clearly at all. You can build hierarchical models and have all these designs, that is analysis, using the outcomes to do that stuff. That is okay, but that is after the design phase. You should invest more heavily in an observational study design than the randomized experiment design, or else you can't believe the answer.

Would you buy a drug that was submitted to FDA and they got a chance to re-randomize and re-randomize and

re-randomize until they got a favorable result? And now they got one regression analysis and they say, oops, that should be the log transform, now I've got approval? You wouldn't believe it. So why should social science have much lower standards? You should have two separate phases.

DR. DUNBAR: Other comments or questions for the panel? Probably only have time for one more after this.

PARTICIPANT: This is in reaction to Don [McLaughlin]'s comparisons of using correlations. The scale score and percentiles and the other metrics are basically -- they are meant to be linear. So I'm not surprised at all what you find early when your correlations are very, very high.

Let me ask you, do you expect them to be different, to begin with?

PARTICIPANT: (Comments not captured by microphone.)

PARTICIPANT: That is because it is much more degraded.

PARTICIPANT: (Comments not captured by microphone.)

PARTICIPANT: Right. So the point is that they are basically different metrics, on the same score linearly and related to each other, and so you don't expect to find differences.

DR. DUNBAR: One more question.

PARTICIPANT: In fact, two comments. One is, in terms of the observational study, there are two differences here from a traditional observational study even. One is that you have the population of schools for some characteristics, so it is just not a grab sum of schools. So to some extent, the set of schools that you are starting off with for some studies is representative.

The other point -- I should have written it down. If you have all of the same schools, you can do studies on the subtypes of schools and find patterns of schools on the descriptive variables that make even further information on the schools you have.

For instance, one thing that occurs to me right away is there are 24 federal programs that give money to schools. What are the patterns that schools are having different numbers of programs? Those types of descriptive things certainly could be done with this data system. Whether you can really get down to cause for a very particular program will depend on whether you can sort out the groups that you could use as contrast groups.

DR. DUNBAR: Thank you all very much. That more or less brings our program almost to a close. I wanted to remind the audience and the committee members that unlike many NRC projects that lead eventually to some kind of a

report that contains conclusions and formal recommendations, this particular project was simply a project to hold a symposium on the use of the school level database. So in a sense, this symposium is the project, and what we have collectively witnessed constitutes a kind of oral report for that project. I think that is a fairly marked departure from typical NRC work.

But you will recall yesterday at the end of the day, I gave you a homework assignment. I want to encourage any of you who may have written anything down or have thought of something for that homework assignment to turn it in to Judy [Koenig] and Lori [Wright]. They will grade the papers and perhaps enter some things into the record.

Don't worry about it. She is trying to put up the list. I read off a list of five more or less consensus-like comments after yesterday's session, and I want to add a few more to that, and also give you an opportunity to chime in on additional things that you may have heard of today.

Those five briefly involved the value of multiple reporting metrics; the importance of systematic documentation of relevant covariates; the fact that quantitative analyses need to be bolstered by very careful observational studies, especially as pertains to issues like program implementation; and also that descriptive

studies based on this data set themselves may have a great deal of value. Understanding what can and can't happen with individual matched data can help us understand our data better when we only have matches at the school level, and the desirability of information about within-school variation.

I heard a number of things today that I thought could be added to this list, so I'll just read these off very quickly, and if there is a serious objection to any of them, please voice it.

Descriptions of the measures, their psychometric characteristics, both the content and performance standards on which they are based are needed to understand the range of inferences that the measures can support, especially if we are talking about using those measures for evaluating programs. That is number six.

Number seven, users of the school level database, for example, evaluators on contract with the Department, need to be clear about the assumptions of their analytical techniques and explicit about justifications for statistical models that they are using to evaluate federal programs. We heard comments from several people who had reviewed reports using this database, and that comment came from many of them.

Number eight, state level databases could be used

to improve or augment the school level database in important ways. Linkages between databases should be considered or perhaps even built in. Based on what we heard about the activities going on within the Department, it might well be that funding streams within the Department that support database development should be analyzed and perhaps coordinated.

Number nine, state and district assessment in data specialists have a wealth of information and experience that the Department could draw on to improve its own efforts in data development.

Number ten, coordinated efforts across the Department are needed to integrate activities of states, for example, in their implementation of NCLB plans, with data development and improvement.

I think we are all tired. If you have anything that you would like to add based on discussions today or yesterday, please share it with Judy Koenig or Lori Wright. I want to thank you all for hanging in there, going with the flow with the weather this morning, and helping us in what I think has been a really stimulating and useful conference on the school level database.

Safe travels.

(Whereupon, the meeting was adjourned.)