

**NATIONAL RESEARCH COUNCIL
THE NATIONAL ACADEMY OF SCIENCES**

**Symposium on:
USE OF SCHOOL-LEVEL DATA
FOR EVALUATING
FEDERAL EDUCATIONAL PROGRAMS**

December 8, 2005

**U.S. Department of Education
400 Maryland Avenue, SW
Washington, D.C.**

Transcribed By:

**CASET Associates
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.

TABLE OF CONTENTS

	<u>Page</u>
Welcome, Introductions, Stuart Elliott, Steve Dunbar	
SESSION I: BACKGROUND ON THE DATABASE	
Moderator, Wendy Yen	
- The School-Level State Assessment Score Database	1
- David Goodwin, Stephanie Stullich	
- Discussing the Challenges of Conducting Program Evaluations - Steve Dunbar	26
SESSION II: THEORETICAL AND METHODOLOGICAL ISSUES	
Moderator, Steve Barnett	
- Estimating Grade-Level Causal Effects	43
- Elizabeth Stuart	
- Designing Gross Productivity Indicators: A Proposal For Connecting Accountability Goals, Data and Analysis - Yeow Meng Thum	62
- Discussion	88
- Can We Infer Causation from Cross-Sectional Data	103
- Michael Scriven	
- Controlling for Student and School Differences: Value-Added and Residual Gains Approaches	123
- Gary Miron	
Discussion and Synthesis Panel	
- Discussion Facilitator - Steve Dunbar	
- Laura Hamilton	149
- Robert Linn	170
- Judith Singer	184
- General Discussion	209

P R O C E E D I N G S

Agenda Item: Welcome, Introductions.

Agenda Item: SESSION I: BACKGROUND ON THE
DATABASE. The School-Level State Assessment Score
Database.

MR. GOODWIN: [In progress.] -- preferable to others, more appropriate than others and, given the limitations of the data, how should we characterize the results, and what kinds of qualifications should we place around the results.

So, that is basically all I have to say and, with that, I am going to turn it over to Stephanie.

MS. STULLICH: Good morning. I think David has covered this topic very well, and he has maybe not left too much for me to say.

So, I might repeat a little bit of the ground that David has covered, but I will try to add some different pieces of information or was of looking at it.

As David discussed, we have two main evaluation offices in the Department of Education, and our roles are, I think, quite different and complementary.

IES [Institute of Education Sciences] is mainly looking at educational interventions, not necessarily federal education programs. They are particularly, I think, looking for opportunities where they can study

educational practices and programs using very high quality, rigorous experimental designs using randomized field trials.

Congress and the administration have a need for information about a wide range of federal programs that may not be possible to evaluate with those kinds of methods, and that is where we come in. We do try to use quasi-experimental designs to the extent that is feasible when looking at those questions.

David mentioned GPRA and the PART, and I will just spell them out for those of you who are not in the beltway and may not use all the same acronyms that we use.

GPRA is the Government Performance and Reports Act. I can't tell you exactly when it was passed, but it was, I think, some time in the 1990s and requires annual reporting on performance indicators for each program in the federal government.

So, each program must develop its own indicators which can be measured regularly. They prefer annually, but some are less frequent than that.

The PART is a different process. PART stands for Program Assessment Rating Tool. It was developed by the Office of Management and Budget, to provide, perhaps in some ways, a more comprehensive assessment of program performance.

In the PART, they look at outputs, outcomes and efficiency. PART is also a verb, for you have been PARTed. It is not necessarily a fun experience, but sooner or later we all go through it.

The PART process results in performance ratings for each program, and those programs can be effective, moderately effective, adequate, ineffective -- you certainly don't want that one -- or results not yet determined.

So far, on the web site, I found some overall statistics across the federal government. About 11 percent of programs have been determined to be effective, 26 percent moderately effective, 21 percent adequate. Only five percent were found to be ineffective, and a fairly large percent, 37 percent, results not demonstrated. Presumably that means yet.

I am just going to skip to the next slide. I thought I would just say a few words about different kinds of assessments that might be used in federal program evaluations, and I think we, at one time or another, have used all three of these.

They are state assessments and independent assessments, and each of them has their strengths and their weaknesses.

NAEP, as we all know, has the advantage of being

consistent across states, consistent over time, and also basically free, not free to the taxpayer, but free in the sense of if your evaluation can use that data, you don't have to pay for it. It is a wonderful on-line data tool.

The disadvantages of NAEP, of course, are that it is not aligned with state standards. So, people may argue that your evaluation is using a measure that is not aligned with what children are expected to know and learn and be able to do in their states. Typically, the data can't be broken out for program versus non-program schools.

State assessments, on the other hand, have kind of the opposite characteristic strengths and weaknesses. They, of course, are aligned with state standards. You can link school-level state assessment results to program status, in federal programs.

To some extent, the data are free, in that you don't have to pay to administer the assessment. Somebody else has already done that. You just need to get your hands on the data, which is sometimes easy and sometimes not.

The data base that Don has assembled for us, Don McLaughlin has assembled for all of us, really, I think made great steps in making at least the school level state assessment data available to us and to you.

The big problems with the state assessment data are, of course, that the state assessments are not

consistent across states and they are often not consistent across time.

There is a lot of activity in the state assessment world, and it seems that state assessments are constantly changing, sometimes in obvious ways, sometimes in less obvious ways.

Sometimes a state will go to a whole new type of assessment, and that is pretty clear, and sometimes they will have the same assessment, but there may be significant changes in how proficiency levels are defined or inclusion and accommodation practices.

So, it is a very fluid field right now, and we keep thinking that we are going to get to a point where state assessments don't change quite so much, but we have been thinking that for quite some time and haven't gotten there yet.

Then third, as David mentioned, independent assessments are, you know, in some ways what you might think as the best of all possible worlds, although in reality they often aren't.

The advantage of an independent assessment is that you could administer the same assessment to all students in your sample and you have it be consistent.

You could also presumably have the control to ensure that that assessment is consistent over time. Since

you are administering the assessment yourself, you would know -- you would be able to link that data to schools that are getting the program or not getting the program.

The big disadvantage is that they really are often prohibitive. One is the cost. It makes an evaluation very expensive if you administer your own assessment.

It is also very burdensome on schools, and they often don't really want to be involved in that kind of study.

It is, of course, like the NAEP, not necessarily aligned with state standards. I think a fourth point is that often you have limited pre-intervention data because, by the time a program is authorized and you go in there with your study and administer your test, the program has often been around for a while.

So, one of the advantages of the state assessments is that, to the extent that they have been administered over time, you can go and get historical data in order to have a longer time frame.

David talked about the origins of our school level state assessment data base. So, I probably don't really need to say anything more about that.

This did start, I would say, with the Improving America's Schools Act, which passed in 1994, and the state assessment requirements were part of Title I Part A.

Prior to that, Title I, Part A, did have assessment requirements but, in general, states had to ensure that students participating in the Title I program were assessed.

So, the big change in 1994 was that the assessments for Title I students had to be the same as assessments for all students, and that is what created this opportunity for evaluation and kind of this whole wave of accountability and assessment that we are in the middle of today.

Characteristics of the data base, I think you all may know this quite well, but we do now have five years of data in the data base, from 1999 to 2003, and it does include nearly all schools and all states with, I think, a few exceptions.

For the most part, it is a data base that has aggregate data at the school level for each tested grade. There are, I think, some areas where we have subgroup data, but predominantly it is aggregate school-level data.

As David mentioned, Don McLaughlin and his team have merged in school characteristics from the CCD [Common Core of Data]. It is all this information about federal program participation, which I think makes the data base easier to use.

Then, initially -- I think this is an important

point for us. Initially the data were collected basically in whatever form they were available.

Don and his team were basically harvesting data from state web sites. So, if the state had scales for us, they took those, and if they had achievement levels, they took those, and if they had something else, like percentiles, or some combination of the above, basically they took whatever they could find.

Although they, I think, predominantly harvested from state web sites, in some cases they went to the states and tried to get more than was actually on the web site, to make it more comprehensive.

We are in the middle of a transition to a new system for collecting this data. It is now being collected by EDEN [Education Data Exchange Network]. It used to be the Performance-Based Data Management Initiative [PBDMI], and I think it is going to be called EDFacts in the future.

So, this data, instead of being a separate stand alone project, is now going to be part of really a massive Department of Education data base, which is collecting all kinds of information about enrollment and program participation and achievement, and a wide range of other things.

As part of that kind of redesign of this data collection effort, one of the decisions that they made was

that, rather than collecting all these different kinds of achievement data, we are just going to collect data for achievement levels only, in other words, the percent of students that are advanced, proficient, basic, et cetera.

One, of course, big advantage of that is having consistent data to the extent that, of course, states define proficiency in different ways and have different assessments. So, it is not consistent in that sense, but it is a consistent type of measure collected for everyone.

A disadvantage, it seems to us, is that, by no longer collecting -- we think that these different types of assessment data, probably some may allow better, stronger, more rigorous evaluation design than others.

So, by losing the collection scales for data, are we losing the ability to do better designs and assessments of states, is a question that we worry about.

So, how do our contractors use school level state assessment data? We typically -- and I think most evaluation contracts that we let have requirements to analyze program outcomes.

Very commonly, the school-level state assessment data is the measure that they are going to be using because it is often the only measure they are able to use within the resources for the contract.

So, our studies will describe and compare

performance at schools participating in particular federal programs, or serving particular populations.

They may examine the achievement characteristics of program schools versus non-program schools as well as trends and, as David mentioned, sometimes identifying schools that are beating the odds, doing better than average, that one, provide a bench mark for what is possible and, two, provide some opportunity perhaps to study what those schools are doing that enable them to achieve those outcomes.

One of the things that puzzles us a bit, and one of the big reasons that we have asked all of you here today, and asked the [National] Academies to do this project, is that basically, although we ask our contractors to use these data, we don't specify exactly how they are to use them.

We are pretty much letting 100 flowers bloom. So, each team develops its own approach within the parameters of the study that they are doing, and the research questions they are trying to address.

It is not entirely clear to us how to make judgments about all these different approaches. Are some better than others. If so, which ones, or are they different because they are appropriately trying to answer different kinds of questions.

Some of the studies are looking at changes in percent proficient or, in some cases, scale scores. Some look at change in relative ranks of schools within the states. Others may look at the difference between predicted achievement and observed achievement.

Another big difference is the issue of whether you can pool data across states in any way. Some contractors, for some of our studies will choose not to do that and do individual analyses within each state in the study.

The disadvantage of that, of course, is that it makes it hard to draw conclusions broadly about the outcomes for the program overall, and frankly, that is the bottom line question we are asked to answer. So, we are hoping to take a big step forward through this meeting and this project.

Just to kind of recap what are the key questions that we are hoping to see answered through this symposium and this panel, as David mentioned, first and above all, What are the most appropriate ways to use the school level state assessment data to evaluate federal programs?

Given the limitations of the data, how should evaluation findings be characterized or with what caveats? We don't want to stretch beyond what is reasonable and appropriate to say with these kinds of data and these kinds

of analyses.

Some kind of sub-questions, one is relating to comparisons across states, and I think for me this is really one of the big ones.

Although state assessments differ, we feel such a strong need to be able to make statements about programs overall, not a 50-state table that shows you what is happening in each of 50 states.

It is pretty hard for a policy maker to know what to do with that kind of output from an evaluation. So, are there acceptable ways to analyze data from assessments that differ across states in order to draw overall conclusions about a program?

Similarly, comparisons over time, state assessments, as I mentioned, are very fluid. When you try to find states that have even three years of consistent assessment data, you end up with a minority of states, and three years isn't very many.

So, to be able to look at time series data for using state assessments, if you can't use a state when their assessment changes, then your pool of states that you can use just goes way down.

A final question is about data needs, ways that this data base could become more useful. Is there a potential improvement to the data that we are currently

collecting, or new data elements that could be added to make it even more useful, basically to you? We would love to know your ideas about that. That is all I have to say, and thank you very much.

[Applause.]

DR. YEN: Our next speaker is Steve Dunbar, who is going to be discussing the challenges of conducting program evaluation.

DR. DUNBAR: I am going to get up there eventually, but we are doing pretty well right now in terms of our time.

So, I just want to say a few words before making any formal comments, and that concerns the fact that we really want, as much as possible, for there to be time for discussion, and this to be as interactive as possible in a large group like this.

As Stuart Elliott explained earlier, we had planned more or less for a fairly -- relatively speaking -- intimate group of researchers and policy makers that would be maybe sitting across the table from each other and really discussing some of these heady issues. We would like, as much as possible, to keep that hope for this larger venue.

Before I talk in more general terms about some issues in evaluating programs, I want to give you folks a

chance to ask any questions, either of David or Stephanie, on background for the data base. Are there factual questions, maybe not for this audience, since many of you have worked with it, concerning the data base, or issues related to its historical development or possibly future development, questions that you would like to direct to our speakers? What if I said I wouldn't go on unless there was a question?

DR. THUM: On point seven, the second bullet, you said that, I believe, the 2003-2004 data would be collected for achievement levels only, and you noted some advantages and disadvantages.

From my own research perspective, I will be very sorry to see that, if that is the way this data base is going, only because I think if you want to try to demonstrate growth on the performance level, there is just not enough information, even when you have four or five, even six time periods.

Strong statistical analysis at the school level of performance, basically distribution, but exactly what it means, is another issue. They are not really anchored on these kinds of measures, progress at a student level.

So, I was very surprised, given the press right now to go with longitudinal data bases, that we are taking this turn here in this very important project.

MS. STULLICH: I guess I would just say that I think David and I both agree with you. I think one of the things that may be useful for this group -- I am not exactly sure how the process is going to work, but if there is a way that there are any kind of requests or recommendations that come up through this process about the needs of researchers and evaluators of the data, I think that would be a helpful thing.

PARTICIPANT: Why was that decision made, to just collect that kind of data?

MS. STULLICH: I think that the decision was made partly due to concerns about cost and complexity of the larger project that this is now a part of, and also I think it was made partly because in the law, No Child Left Behind; the focus is so much on achievement levels and on students achieving proficiency on state assessments, that it was felt that this approach was very consistent with the law.

DR. DUNBAR: Laws and policies for implementation change, as we know. A question over here?

DR. JOHNSON: I am Kirk Johnson from the Heritage Foundation. I would like to underscore what the previous two questioners have said regarding the loss of scale scores.

I think it is very important, because so many

states, at least anecdotally, were changing what constitutes proficient from a year to year to year basis.

Even if you had achievement levels, because the line for achievement seems to be moving, it would be much more -- it was much easier to analyze scale scores that might not change as much, or at least the actual scale itself might not change, even though the proficiency levels seemed to be moving targets.

So, I would, too, view it a great loss if we didn't have as much information as possible out of the school level data bases.

DR. DUNBAR: We will be talking more about scaled scores and things as the day goes on.

DR. MCLAUGHLIN: I have a fair amount to talk about in a half hour tomorrow, and I have some things to say about the history of the data base. It started before 1998. I wonder if it would be reasonable, if you have extra time, if I could present one of the slides.

DR. DUNBAR: Why don't you try to talk us through what you want to say about that?

DR. MCLAUGHLIN: Fine. Actually, it started, these current efforts of an integrated data base of state assessment scores started in 1994, with a NAEP secondary analysis grant.

Its purpose was to determine whether Westat had

done a good job of substitute schools in need. So, AIR had a subcontract to go out and get the data on state assessments, so we could compare state assessment scores on the refusing schools in NAEP versus the substitute schools.

The outcome of that -- well, the primary outcome was that it looked very good for NAEP. There was virtually no difference, certainly no significant difference between the refusing and substitute schools.

The background for that was this data base. At that time we had 23 states, and we held onto those, the school-level state assessment data.

In 1997, Mike Roth at NCES [National Center for Education Statistics] supported our efforts to link those data to the schools in the staffing survey. In fact, it is tab 13 in the background data, for those of you who have the background readings.

We were able to show some value in having these data, that we could merge them with a data base that didn't have an achievement measure.

Then, in 1999, with the new NAEP contract, we had a project to help the states based on NAEP. Of course, that project, we collected the data from virtually all of the state assessments, modeled virtually all the states, and sent out reports to each state on the relationship of their own assessment results to the NAEP results for 1998,

and continued that since.

It was, as was pointed out, in 1999 that I had a discussion with Alan Ginsberg that led to funding from PPSS to do a 12-state trial to see whether we could collect these data to be used in federal evaluations, in fact. We did it for 49 states. It took us a while to get Iowa.

In 2002, the successor of PES, FAS, decided to drop the funding for our development in favor of developing the PBDMI.

Fortunately -- I consider it fortunate -- we had the NAEP state analysis project, and NAEP was very interested in continuing to prepare NAEP state assessment results. So, we continued to collect these data.

The 2003 data were collected, funded by NAEP, and we are continuing to collect the data. The 2003-2004 data and the 2004-2005 data are available.

At present, Victor Bandeira de Mello at AIR is collecting those data. I think he may be looking into them, and we hope to have the -- we have 30 -- he has now 30 states, at least, for 2003-2004 and 2004-2005.

He is collecting whatever data are on the web from the states. So, in the data base, if the state is giving scale scores or percentile rank or another measure, it will be on what we call the NLSLSASD, and those data for the two most recent years should be available some time

next spring.

I should point out that I think it is a false consistency to say that it is just to collect the data on the percent achieving some level of proficiency because, as someone mentioned, they vary all over the place, and the scale scores are probably a much more stable measure. On the other hand, as I will talk about tomorrow, they are all highly correlated.

PARTICIPANT: Don, maybe you could hold that for a minute, but I think one of the concerns we have, I know that you are collecting 2003-2004 and 2004-2005 data, but I think one of our concerns was whether you would be able to continue to find the funding to support that on an ongoing basis.

DR. MCLAUGHLIN: Our current funding is from NAEP, and that is a five-year project, the NAEP state analysis project.

I believe the plans are to do it for 2005-2006 and 2006-2007, but I will tell you, we get funding year by year. I don't know for sure, but that is sort of what we have put in our proposal.

PARTICIPANT: That makes me want to ask a more general question, and this may or may not be the right place to ask it.

I am sort of curious about, for example, is there

sort of a panel or a group that makes policy decisions about the data base, that for example decides, we are going to uniformly move to a common percent proficient metric, just as an example?

What kind of mechanisms or organizations does the department have in place for developing and maintaining this as a data base?

I have some impression, from looking at some documents and looking at the history, that it is almost ad hoc in some ways, and then it has grown to a point where there is a sense of its utility.

So, I am asking, are there administrative mechanisms in place for its continued support? How does it work?

MR. GOODWIN: As Stephanie mentioned, the responsibility for supporting the development of this data base no longer rests with our office, but with another group in the department.

We have periodically commented to them about the desirability of having more detailed scale score data rather than just proficiency levels.

Basically, it is a trade off between what are a series of administrative records, and the efficiency of collecting those in a manner that is consistent with No Child Left Behind, and the requirements of a research data

base. You could imagine just how valued the latter might be, relative to the other priorities.

My guess is that there is going to have to be some effort to use those data in the current manner in which it is being collected, before people will realize the limitations of that data, in terms of what they want to get out of it, and then perhaps respond to that by going back to states and collecting more detailed data.

Right now, I don't know that there is a formal structure for making decisions about what is and what is not in the data. It is part of a much larger data collection effort that involves all kind of information about programs that historically states have reported to the department.

Each separate program has reported its separate performance data to the department, and now this has been integrated into one major omnibus data collection. So, this is just one item, a key item, but one item in a larger data collection.

MS. STULLICH: If I could just add, Steve, the way that you frame the question is very interesting, as if it seemed like there was this sort of ad hoc -- in fact, the impetus for this new sort of mega data base was exactly that concern, that there were all of these different offices collecting different data throughout the

department. So, let's bring it all into one giant, consistent, coordinated data base.

I guess there is a little bit of a lesson, perhaps, of be careful what you wish for. When you do that, it seems that maybe one of the side effects that comes along with it is that, when you have an individual project in an individual office, that office may care a lot about some of the details that to us they may seem central, but to somebody else they may seem like details that get lost in the pursuit of this overall objective.

PARTICIPANT: It sounds to me as though, in the past, this has been a rather passive involvement of states. You went out, you saw what states had on their web sites, you took it, without very much involvement of the state.

It sounds to me, though, as now the state is becoming involved, and perhaps in two ways specifically. First, it is providing information about what is going on in schools in the state.

So, the first question is, How is that solicited and screened? The second, if the theory of involvement is the possibility of negotiation with the state about the nature of the state's submission in terms of achievement there, I wonder if that opens up the possibilities for developing newer statistics than pursuits in achievement levels.

MR. GOODWIN: I am not really an expert on exactly the process that is unfolding in terms of collecting data from the states.

I know there have been elaborate discussions about specifying the definitions of the data items and, wherever possible, to have consistent data items across programs. So, you know, a high poverty student for program X means a high poverty student for program Y and so forth.

I would say that so far it has clearly been transformed from a passive kind of harvesting data from the web to something where the states have to actively supply data in a consistent format that is generally specified by the department.

This is a burden. Frankly, so far, I understand that they haven't had total cooperation from the states, particularly on the achievement data.

I think for the 2003-2004 year they probably got half the data from the states. Probably Don [McLaughlin], in the manner he operates, probably has most of the data from the states.

The department is going for a kind of consistent format. To lay on top of that, it seems to me, at this point the requirement that they also supply scale scores is likely to mean it will make it more difficult, even, to get complete coverage across the states, at least at this

point.

DR. MCLAUGHLIN: I don't know if it is obvious to anyone, but there are two data collections, two versions of this data base going on now.

Up until 2002, it was one data base, and now the department, it seems, is continuing to develop the data base in one direction and AIR is -- I guess I should say NAEP is collecting the data in a somewhat different way.

Yes, we are more passive but, in fact, one of the rules throughout has been contact with state assessment directors.

More recently, we have been contacting the NAEP coordinators, someone we can work with and get the data. It had been a kind of a dialogue with purpose since 1999, to give back to each state some information about the relationship of their state assessments and needs.

They may not care about that but, by extrapolating from that, they can compare their state assessments with state assessments in other states by how they both compare to NAEP.

So, we considered it a dialogue between the fed - - me -- and the states, not just the states are providing the data. We have been very passive in the sense of what the states have to provide for what is on schooldata.org.

MR. CHAPLIN: Duncan Chaplin, Urban Institute. I

just wanted to join the band wagon for scale scores here. The one issue that I am wondering if anybody has thought about is, if you are looking at gaps and how they change over time.

It seems like there are a lot of policy interests in that, because of No Child Left Behind. My sense is that proficiency scores might end up telling very different stories than the scale story in that particular situation, and that might be the major reason.

So, as everybody is starting out at zero percent proficiency, as you move up toward 50 percent, you would expect gaps to grow. Above 50 percent, as you move toward 100 and everybody gets stuck at 100, you see the gaps shrink, even in the absence of any change in the gap for scale scores.

So, it seems like No Child Left Behind might have sort of a contradiction built into it where it says, we have got to look at proficiency and we have got to look at gaps.

Well, if the gaps are just going to be driven by this mathematical issue, that seems like a problem. I don't know if anybody has thought about that.

DR. DUNBAR: There are good scale scores and there are bad scale scores. I think we want to be real careful when we jump on a particular metrics band wagon.

How that metric was developed is something we can't necessarily make assumptions about. One more question.

MR. DUNCAN: Well, good and bad scale scores, yes, but the advantage of adding the scale scores is that you get at least a second shot at a perspective on the situation.

DR. DUNBAR: Absolutely.

PARTICIPANT: I was on one of the teams that did the evaluation of the original NAEP. The thing that we were not allowed to discuss at all was the political panel's decision on naming the cuts.

It was -- since we took the time to go through every single item on every single test -- it was apparent then that proficiency was a ludicrous exaggeration of the achievement level that was represented by that score.

We have sort of gotten used to using the term proficiency now, but we shouldn't let it slide like that. It has now become sort of a band wagon.

The states are now in charge, to a large extent that they weren't before, and they of course like the White House suggestion to get more people proficient. Well, the easy way to do that is to redefine proficient. So, the scale scores, however bad, might keep it slightly in perspective.

DR. DUNBAR: I agree with you. I think it may be

the lure of a single metric that we should be wary of.

DR. YEN: Steve Dunbar now is going to be discussing the challenges of conducting program evaluations.

Agenda Item: Discussing the Challenges of Conducting Program Evaluations.

DR. DUNBAR: Now I might be able to fill the allotted time. Thank you very much for those questions. I am going to try to just lend a little bit of perspective on what we hope to accomplish in the remaining part of the workshop.

Our purpose in organizing this the way we did was, first, we wanted there to be at least a time for the department to describe what is and how it came to be.

That is the reason we asked David [Goodwin] and Stephanie [Stullich] to really open up this workshop, and the reason I wanted to make sure that there were no questions sort of factual about the nature of the data base, the beast that we have got.

Our perspective on this as a steering committee for this workshop was to try to examine some of the underlying purposes of having such a data base, given the importance of it in doing federal program evaluation.

To do that, we more or less posed some important questions about statistical modeling, about measuring,

about policy and practice in education.

Our early discussions, when we considered the data base and its limitations, to be perfectly honest with you, we scratched our heads a lot and thought, gosh, with all these limitations, what possibly can we do.

We were very concerned that we would have a two-day workshop where, at the end, we would say, well, enough of this, we can't do much of anything with it.

We decided that simply wasn't an option. That wouldn't be a very responsible or responsive approach to the department's needs or the imperative for evaluating federal programs.

So, not from the very beginning, but after the first morning of our steering committee meeting, we decided that what we needed to do was frame important methodological questions that involved statistical methodology and educational measurement, in such a way that we could then address what we believed to be the important agenda of this workshop.

How can we, number one, use the data base as it exists for the purposes that the department wants to use it? Number two, what concrete suggestions can we make to the department for improving a data base, given the imperatives associated with federal program evaluation, and given the fact that more and more states are bringing

various kinds of data pertaining to education policy and practice into an electronic format, so that, perhaps things that we really haven't conceived of at this point could, in fact, become part of not just a data base, but some kind of information warehouse, and maybe that is a term you all can hang onto for a little bit?

So, I want to talk about some basic questions regarding those issues, and give you an overview of the rest of the workshop, so that you have a sort of an advance organizer, so to speak.

If we are asking questions about the challenges of conducting federal program evaluations, I think there are some basics that we have to get out on the table.

Some of these questions will be things that we simply dismiss. Others we will spend a lot of time talking about.

I think, number one on the list in an evaluation mind set is specifically what do we value. That is often a given in federal program evaluations.

David described it very clearly, that we want to see gains in student achievement, we want to see closing of achievement gaps, and those are the big picture questions, and values in this sense are more or less taken as a given.

On that very gross level, we all share common values. When programs are implemented, however, there are

lots of value decisions that are made in the process of implementation, whether it involves delivery of programs, services, interventions, or whether it involves the kinds of things we do to measure the outcomes of the interventions.

There are values inherent in approaches that we take, and I will give you a few examples of those in a minute.

Another fundamental question, of course, is, once we have decided what we value, how do we measure it. Developing measures is something that I spend, in fact, most of my time doing and, when people are concerned about standardized tests, they are really talking completely about not what is valued, but how, in fact, you have assembled materials in order to measure it. That is not a trivial matter at all.

A reading test in one state may look like a reading test in another state, and yet may be quite different and tell you quite different things about what children can do. So, we are going to talk about how we measure what we are after, to some extent, during the course of today and tomorrow.

We also want to know, in the context of an evaluation, when and how we can be sure that whatever it is we are measuring has changed.

We might think, just by looking at it at time one and looking at whatever it is at time two is going to tell us that.

We are going to see, in fact, right away this morning that that is not quite as simple a matter as it sounds.

Then the question dealing with data warehousing really is where do we document all of this, because it is, in fact, many times programs need to be evaluated after implementation, and the evaluations have to take place sometimes on an ad hoc basis because of the AYP [Adequate Yearly Progress] programs are implemented.

Without there being a very clear idea of where information relative to program implementation, as well as outcomes, is documented, well, that is just something to keep in mind.

Every time I come to Washington I try to find a way to fit this graph in, which is our achievement trends in Iowa over the last 40 years or so.

It is not quite up to date now, but these are ITBS [Iowa Tests of Basic Skills] scores in grades three to eight in scale score units that were established, oh, as a reference back in 1965 as the scale score metric as it was defined, and then scores over the new forms of the Iowa Tests of Basic Skills that were adapted to that metric.

Maintaining such a system of scaled scores is quite a trick, and to do so we make quite a few decisions about some of the questions that I had on my previous slide.

For example, what do we value? We have decided more or less ad hoc in Iowa that a measure of basic skills in core achievement areas in the school curriculum is about all that we could do in a state that emphasizes complete local control.

We don't have a state department of education that monitors or decides what individual school districts test. We have really a system of schools, rather than a school system, in our state.

So, a very general measure of achievement was devised and has been used up until a few years ago on a voluntary basis by schools throughout the state.

Now, these achievement trends are an incredible asset for us as a state. In fact, their existence is probably why Iowa was able to get an NCLB plan approved that didn't really have a formal set of content standards statewide for purposes of implementation.

The trade off was this. We get these kinds of trends by probably making some compromises in terms of the nature of the measure.

Our measure is much more general than some of the

very highly criterion-referenced statewide assessments that are used in other states.

So, you see there are trade offs. The question of what we value versus how we measure it, those questions are fairly intricately entwined.

How do we determine that changes happen? Well, these are more or less census data. So, I look at the trends and I see that things go up and down much the way that achievement seems to go up and down nationally.

The drop that you see that began in about 1965 at grade eight was the great score decline in college achievement test scores in the early to middle 1970s. So, we have seen some validation of these trends over time.

Those key questions, what we value, how we measure it, when are we sure it has changed and where do we document it, are all sort of critical components of the question that we are addressing today, how do we make use of a large school-level data base for purposes of federal program evaluation.

Now, what you are going to hear from speakers during the course of this workshop concerns those components.

First of all, this morning, after our break and after any additional discussion that we have for these presentations, we will be hearing people talk about how to

define change, both statistically and conceptually.

Questions like, what are the sampling conditions, units, outcomes of the program. When is a given statistical model appropriate for describing change?

How do data structures, cross sectional and longitudinal, influence inferences that we might want to make about change, will be subjects that we will be hearing about in what we viewed, in organizing the workshop, as addressing some of the important statistical foundations for program evaluations?

I think it is tomorrow that we will be talking a little bit more about measures of change. How the measure was designed I alluded to a little bit, and you may address that question implicitly in a lot of the discussion that we have.

Bob Linn is going to talk about what has been done to make sure the measure hasn't changed, whether it be a metric for reporting such as percent proficient, whether it be scaled scores and whether they are linked across multiple forms of a given state's assessment.

Then we are going to hear some new ideas about documenting information about the measures that we have, and what possibilities there are for incorporating new kinds of information into perhaps an expanded data base.

Then we thought it was important, after

considering issues related to statistical questions, modeling and so on, related to designing and linking of measures, that we hear from the people in the trenches, so to speak, from people who are in the position, for example, of responding to requests at the federal level for data that can be assembled.

So, we are going to ask some questions of these folks about what really does it mean for the department to say, we don't want just percent proficient after all. We really need scaled scores, percentile rank, we need to know percent free and reduced lunch. We need to know everything possible that you have available to us and that you can easily get into a format that we can use.

I posed a question to David just in this early discussion, because I think it may be important to talk about what mechanisms are needed to maintain and support a larger expanded data base if, indeed, it seems to be the direction our conversation goes, that what we have now is a good start and demonstrates potential value for program evaluations, here is what we need to really make the thing work right.

Then, in our closing discussion, we are going to examine not just practical issues regarding implementation, but some of the ethical implications of creating, at the federal level, a data base that pertains to local

jurisdictions.

I think that is where we might really get into some interesting issues about the range of possibilities in terms of data warehousing, for example, and issues related to the existence of such data in a central location.

That is more or less an overview of our thinking behind the design of this workshop, and what we hope to accomplish during presentations and discussion.

We are now still ahead of schedule and have time for you to raise questions from the floor, either about the remarks I made, or continue discussion about the data base itself. Wendy, I will turn it back to you and let you entertain questions.

DR. YEN: Are there more questions related to the presentations this morning?

PARTICIPANT: This is a little off topic, but I guess I would be interested in knowing what story you tell about the -- I am not quite sure how to describe that pattern of data, but it doesn't seem to follow a cohort pattern.

Based on your noting that it followed the SAT dip, could you comment about that, or tell us what you think is happening? You have those ups and downs all along the line.

DR. DUNBAR: Well, one of the things that -- you

are right, it is a little bit off topic. I will answer your question, but I don't want to get the whole discussion off track.

What those trends showed us, you are right, you see some ups and downs at grade eight that you don't see, for example, at grade three.

Now, our tests measure different skills. As you go up the grade levels, the skill set becomes a little bit more complex in terms of cognitive components, a little bit more informational, factual in the early grades.

One of the things we observe, not just in those trends which are for composite scores, but particularly in the area of early reading is, for the very first time beginning in 1996 or 1997, we began to see in state test scores in reading comprehension, third grade scores going down, and we had never seen that in the history of Iowa tests up until that time. Those plots started in 1955 and went through about 1995.

We saw 40 years of relatively steady increases in early reading skills and, beginning in 1995, they started to take a downturn.

That got the attention of state policy makers more than anything else, much more, in fact, than the score decline in grade eight did in the 1960s, because that seemed to be happening everywhere.

It led to class size reduction legislation, and it led to our state taking really a very careful look at changing school population that was coming in early. So, those are some of the inferences or interpretations of what you observed as not a consistent cohort pattern across those grades. Now, for the topic at hand.

PARTICIPANT: I was a little curious or taken aback by the fact that there are probably going to be at least two data bases as Don talks about it, the NAEP version and the ED Facts version.

It seems likely that there could be many versions of the data bases, as other people and groups take the core set of data and want to add their own things.

I wonder to what extent does the department see this as a positive thing, and might there be efforts to sort of promote the commonality among various versions of the school level data?

MR. GOODWIN: I was actually a little surprised to learn that the department was sponsoring two parallel data collections, but I guess that is not the first time that has happened.

I hoped that would not come out in the first half hour of discussion, but seeing that it has, maybe we should concentrate on the virtues of having more than one data source.

Perhaps Don is going to collect more detailed scale score data, for example. That will be useful to researchers in a way that perhaps the other data set may have some limitations.

The data are out there. It is in the public domain and there are going to be a lot of people interested in it, and if they don't use the existing data sets that the Department [of Education] or AIR or NAEP are collecting, they might just as easily go out and compile it themselves.

MS. STULLICH: I think having one consistent unified data collection seems like the best way to go about this, but if that one system doesn't have everything that people need, then I think it is sort of not surprising that it has come about that we have a second system now that is trying to basically fill that gap and meet that need.

Now, maybe they will come back together at some point. Who knows what the future may bring, and this conversation may play a role in what happens next.

PARTICIPANT: Early in the opening remarks you spoke about the data base and described that some of the characteristics of it are changing by the use of PBDMI and EDEN.

Could you describe briefly what some of those requirements are to the states as you transition or move

toward what -- I don't think you used the word improvement, but what the outcome would be for what states are supposed to provide that leads to this improvement in the data base through this process of EDEN?

MS. STULLICH: I am not sure I understand the question.

PARTICIPANT: What is it that you are requiring states to do through the PBDMI process and EDEN, that lead toward this data base?

MR. GOODWIN: I am not an expert in what has gone on, but there has been a lot of discussion about developing common data elements across programs to describe very often the characteristics of students who are participating in programs.

In the past the department has, I don't know, 150 different K-12 programs, and they have all had requirements to report data to their program offices in the department.

They have done it and all the requirements have all been developed independent of one another to meet the unique needs and interests of those programs.

Very often there has been common information about the number of participants, the number of schools getting grants, the characteristics of the participants.

Each of these different, separate, performance reporting entities have had their own definitions, if they

have had definitions at all, and I think a lot of work has gone into moving closer to a set of consistent kind of definitions than was the case in the past. So, that is one thing.

Presumably that will reduce the burden on states, but also have some consistency, so that high poverty means the same thing when it is reported in the case of vocational education as when it is reported in the case of some elementary and secondary education program. So, that is certainly one kind of improvement, I believe, that is likely to occur as a result of this overall effort.

DR. YEN: I think we have time for one last question.

DR. MCLAUGHLIN: Just a comment, one of the sorts of missions that we have is to make these data available. So, Victor [Bandeira de Mello] and his staff [at AIR] clean up the data, and they won't be out until next spring, but the aim is to make it available for whoever wants to do analyses using these data.

We understand that some of the data there on the data base will be different from what is on the state web sites, because we found the states would change.

They might have something that they report out on their web site in June and then something else in August, and we take what is there.

Our aim has been trying to make these data available, and others have used these data. One of the things that we said to the states that we hoped would happen and it hasn't really happened is that we would have these data available so that people wouldn't have to call the state.

I understand from Victor that the states are telling us they are still getting a lot of calls for the data, but I think that would be the advantage of having one data collection, and one warehouse that is completely publicly available.

I should say, one of the arguments against making it publicly available is the suppression of small samples. That is one of the things, we have taken care to suppress the small sample data. That is one of the problems with the data base, especially for a study in minority gaps.

PARTICIPANT: Let me just follow up briefly. It sounds in general like the ability to get data from states has been improving over time, but I was wondering if there were any particular states that have cut back a little bit, and what sort of issues have come up. I mean, one of them you just mentioned, but if there are any others, it would be interesting to know.

DR. MCLAUGHLIN: Actually, I can't really answer that. Victor, who couldn't make it here, is now collecting

the data, and he has said there is some problem of states being tired, and states having a priority list that is very long in terms of requests for the data.

PARTICIPANT: I think that is one of the things we will want to hear about especially tomorrow, is all of the wide range of burdens that are placed on states right now, just for data collection and warehousing within their own jurisdictions, let alone for federal purposes. That is a real important topic of conversation.

DR. YEN: That is the end of this session. We are starting again at 10:30, and see you back here then.

[Brief recess.]

**Agenda Item: SESSION II: THEORETICAL AND
METHODOLOGICAL ISSUES. Estimating School-Level Causal
Effects Using the SSASD.**

DR. STUART: I was just saying, it is nice to be here. It is only across the street from my office, but I am going to try to stay over here and not feel the draw, the pull, of my office itself.

I currently work at Mathematica Policy Research. I have been there just over a year. I did my graduate degree in statistics at Harvard, and worked with Don Rubin, who will be here tomorrow.

So, most of my graduate school work was on causal inference, and that is really what I am going to be talking

about today.

My dissertation was on a high school drop out prevention program, and some new methods for analyzing that data.

I have taught one-day courses on causal inference with Don. So, this is really a topic that I have thought a lot about over the last six or so years. At Mathematica, I am now getting a chance to really see some of the more practical issues involved, and I think today is sort of a nice combination of these two pieces.

I am going to provide sort of an overview. I thought as one of the first speakers it would be helpful to sort of take a step back and think about what is the sort of theoretical framework that we are talking about here.

So, these are some of the motivating questions that people interested in using this data base might be trying to answer.

Does increasing funding to school libraries improve literacy? Does the Reading Excellence Act improve test scores in high poverty schools? What is the impact of accountability policies and practices on Title I schools? How does expanding school choice options affect test scores?

In fact, these are all addressed in some of the papers that were sent to me, at least, to look at. So,

these are all causal questions.

What I mean by that is that they are explicitly trying to state what is the effect of some intervention, some intervention that we could apply or withhold.

So, this is distinguished from descriptive studies that are just looking at trends or sort of analyzing how data is changing.

In these cases, we are really interested in saying, what is the effect of this particular intervention. So, today I will talk about how to answer these questions in a sort of broad way, and really stress a couple of common problems that are encountered.

There are two main points that I hope you can come away with. First is the need to clearly define the question of interest. Second is the importance of sort of comparing apples and applies. It sounds logical. Sometimes it is harder to do in practice.

Before I move into some of those issues, I want to sort of take a step back and just think, What are we really trying to estimate?

So, we want to think about how do we actually define causal effects. A causal effect is inherently a comparison of potential outcomes on a common set of units.

So, we might have a potential outcome under treatment, some intervention of interest, for example,

school tests scores with extra library funding, and a potential outcome under control, which is these test scores without the library funding.

So, we would really like to compare these two potential outcomes on each unit. So, this is sort of the truth that we can imagine having.

We have a series of units. Each row is a different unit. These might be schools. For each of these units, we observe some covariates. These are variables that are not affected by the treatment, ideally measured before the treatment would even be applied.

Then we have two potential outcomes for each of these schools. The impact for each school or each unit is a comparison of these potential outcomes, for example, the difference, maybe the ratio, some comparison of Y_0 and Y_1 .

This is the ideal world. The problem is that we never actually get to observe all this data. We, instead, observe something like this, where half the potential outcomes are missing.

Each unit either gets treatment or gets control. We therefore can only see one potential outcome for each unit. Therefore, all of the individual impacts are unobserved as well. So, thinking about causal inference as a missing data problem, where half the data of interest is missing.

Before I go on again to talk more about how to actually estimate these causal effects, I think it is important to think about the SSASD in particular, and what are the kinds of treatments and units that we can think about.

In general, the main theme is it has to be schools. The treatments have to be school-level interventions. There may be a district-level intervention, some unit higher than the school itself.

It cannot be with this data set, because it does not have individual student data. We can't look at the effects of, for example, pulling students out of their classroom and giving them more instruction in reading. We can't look at individual level interventions.

The potential outcomes of interest are probably going to be test scores. That is what is available in the data set. Again, the units are going to be schools or maybe higher-level districts.

The one thing that is really important here is to beware of what is known as the ecological fallacy. What this says is that the trends that are observed at, in this case the school level, can't be assumed to hold at a lower level, such as students.

One sort of hypothetical example is that we might observe a positive association at a state level between the

percentage of foreign born students and reading test scores. There may be more foreign born students in states with higher test scores.

We can't assume that that means that foreign born students, as individuals, are more likely to have higher test scores. In fact, it is likely the opposite, because fewer of them will be native English speakers.

In fact, what is really going on is that foreign born students are more likely to move to states with high test scores. So, you can't assume that these relationships at the state level hold at the student level. I think that is important to keep in mind as these analyses are done.

So, how do we learn about causal effects? I have sort of given a three-minute introduction to the finding of causal effects, and now I will move into estimating them.

There are three common approaches that I am going to talk about today. The first is randomized experiments, the second are sort of pre, post or before, after studies. Then, observational studies and comparison group designs.

Randomized experiments. Probably everybody here has sort of heard, randomized experiments are generally considered to be the 'gold standard' for causal inference.

A lot of that is due to the fact that, because units are assigned to treatment or controlled randomly, those groups, the treated group and the control group, will

be equivalent on all of the baseline covariates.

So, ahead of time, before the treatment was applied, the treated and control groups would be statistically equivalent. There would be no systematic differences.

That means that any difference seen in the outcome can be attributed to the intervention and not to any of these pre-existing differences.

I am not going to talk about randomized experiments much in the context of the SSASD, but I do want to point out that I think there are a few places where it can help.

This might be in defining a sample frame, again, since it has data on nearly every public school in the country.

It might include how to select schools for inclusion in the study, or reducing data collection. So, if organizations can identify the schools in the data set, they can use that data set instead of having to collect the data themselves.

The problem, though, is that randomized experiments are sometimes infeasible. So, what can we do instead?

A common approach is to do pre-post designs, where test scores are compared before and after some

intervention. We might compare test scores in the year before an intervention is applied, and test scores in the year after.

I am going to use an illustrative example to explain why this can be very dangerous, and we need to be very careful about making causal statements in this setting.

Some of you might be familiar with Lord's paradox. This is an example from Frederick Lord, who was a psychologist in the middle 20th Century.

It is actually often used in discussions of gain scores versus regression analysis. Should we look at a gain score or should we look at regression?

I am going to use it instead; I think it is a really good example of why it is important to think carefully about causal effects.

So, here is the scenario that Lord laid out, now 40-some years ago. A large university is interested in investigating the effects on the students of the diet provided in the university dining halls, and any sex differences in those effects.

Various types of data are gathered. In particular, the weight of each student at the time of his arrival in September and his weight the following June are recorded.

I am basing my discussion on a paper by [Paul] Holland and [Donald] Rubin. So, I am basing my discussion on a discussion of Lord's paradox, and actually, I do have copies of the Holland and Rubin paper. If anyone wants them, you can see me after.

So, the setting is something like this. The x-axis shows September weight. The y-axis shows June weight. Among the females and among the males the distributions are the same in September and in June. The average weight is the same, 120 for women, 180 for men, the variances are the same. There is no difference in the distributions at the two points in time.

So, Lord posited two contradictory statisticians that came along and came to two very different answers about this differential effect of the university diet.

Statistician one said, well, the women didn't gain any weight, the men didn't gain any weight. Therefore, there must not be any differential effect of the university diet. Zero minus zero equals zero.

Statistician two comes along and says, well, I am going to try this regression tool. I am going to regress June weight onto September weight. Can we predict June weight from the September weight?

If we do that, we essentially see that, for two individuals with the same September weight, a man will

weigh more in June than will a woman, even if they had the same initial weight. Therefore, the diet must have a larger effect on the men.

So, these are two very different answers. Who is right? I won't make you actually vote, but people who have read my paper sort of know the answer. I will just let you think about it for a minute.

So, my answer might be what is expected from maybe a statistician or someone who doesn't really want to answer the question: "It depends."

In fact, we can get either of these answers depending on what we are willing to assume. So, let's think about it in the framework that I laid out in the beginning.

The units are students. Covariates that we observe are sex and September weight. These are things measured before the university diet.

The potential outcomes of interest are the June weight under treatment and under control. The treatment, it is pretty clear, is the university diet. We are interested in the effect of the university diet?

What is the control? There is no control mentioned. We have no idea what this comparison of the diet is with.

So, the data look something like this where, for each student, we are missing the potential outcome under

control. We don't observe the potential outcome under control for anyone and, in fact, it is even undefined for everyone.

So, really, by filling in any numbers in that column with different assumptions about what that control weight is, we could come to any answer about what the differential effect on men and women is of the university diet.

For example, statistician one is essentially assuming that the June weight and the control is equal to the September weight.

Statistician two, instead, is saying, no, the June weight under control is a linear function of the September weight, and in fact it is parallel for the two groups, for men and women.

So, it looks something like this, where the difference between the two parallel lines is the differential effect.

Really, either of these statisticians could be right. There is no information in the data about which statistician to believe.

It would come down to a substantive discussion of which assumptions are more reasonable in the settings of interest.

I sort of thought of a third example which

relates to the freshman 15, and weights under control are September weight plus 15 pounds.

So, that would be sort of a third approach that could be taken and, again, any of these might be reasonable, but it would come down to a substantive question of which assumptions are we interested in.

To tie this back to the SSASD, this paradox is relevant for many educational studies. We can imagine a question such as, does No Child Left Behind have a differential effect on black students versus white students?

In that case, the question is, What is No Child Left Behind being compared with? You can't just answer this question without having some sense of the comparison.

So, without a comparison group or a control group of some type, you have to be very careful about stating causal effects without clearly stating what those assumptions are.

I will move now into observational studies, sort of pre-post. I have illustrated some of the problems with just looking at before or after studies.

So, now a more common situation is where you have actually a comparison group. You do observe that some schools in this case received treatment and others received control.

The key point in observational studies is that we need to control for differences in the observed covariates. This might be done through regression adjustment or matching methods, where we try to find schools that look as similar as possible before the intervention was applied.

One note that I will come back to is that, in these cases, it is very important to include as many covariates as possible, because with observational data there still might be unobserved differences between the groups.

If we include a lot of covariates, we would believe that there are fewer other potentially unobserved differences.

Now I will sort of talk about the dangers of comparing apples and oranges. Again, observational studies often fit a model of the outcome conditional on covariates in the full treated and control groups.

For example, within a state we might compare schools that got some intervention with all other schools in that state.

The problem is that, if the treated and comparison groups are very different from one another on the background covariates, the estimates will rely a lot on model extrapolations and the model assumptions.

In fact, I have some references at the bottom

that illustrate the dangers, or the incorrect inferences, that can be made because of this extrapolation.

I will illustrate the problem here. We can imagine there are some pretest score on the x-axis, and the outcome, post-test score, on the y-axis.

We have a treated group down here, and a control group up here. Again, as I have talked about, we are interested in imputing the missing potential outcomes.

So, for the treated group, we would like to impute or fill in what their outcomes would have been under control.

So, to do that, we use the control group. We might fit some model of the outcome under control, Y_0 , given X . We might fit a straight line and get very good diagnostics -- yes, this is a great fit, this regression works really well for the control group.

It might be that the true model actually is very non-linear or even moderately non-linear among four values of X that the treated group has.

The problem is that we don't observe any control units down here. So, we just don't even know what the true model is in that space.

Again, the regression diagnostics we look at might look fine for the axis for the control group but not for the axis for the treated group, but we just don't even

know that.

A better situation is something like this, where the treated and control group have overlapped on the distribution of the covariates so, again, they look at similar as possible ahead of time. So, any models that are fitted, are fitted in this comment space.

I don't have a lot of time to go into sort of the mechanisms of how to do that, but there is a statistical literature that has been growing on matching methods.

So, how do you implement this? How do you choose these schools that, ahead of time, look at similar as possible?

Some tools that are used for this are propensity scores, exact matching on covariates, Mahalanobis metric matching. Really, again, the idea is to just find schools that look at similar as possible ahead of time.

I will just quickly mention two particular benefits. One, again, we are comparing similar schools, but also the outcome variable is not used in the matching. So, we choose these comparison schools without even looking at the outcome.

This means that they are being selected without any potential for biasing the results in order to get sort of the desired results.

Just like a randomized experiment doesn't use the

outcome in design, the design of an observational study should not use the outcome in a design, and matching methods can accomplish that.

What is really sort of the best approach -- this is, again, sort of moving slightly ahead -- is to really combine the benefits of regression and matching.

Regression can work very well when the groups look like this, sort of just small differences between the groups. With matching, you can get the matched samples, and then run regressions on the matched samples.

What about the SSASD and the studies that currently use this data set? Are they comparing comparable groups?

I would argue that there could be steps to do better comparisons of comparable groups, or at least diagnosing it, at least, and showing how similar the groups are.

One point I will make is I think all the comparisons that I looked at were done within states, and then aggregated across states, but all the comparisons were within states.

This [table] summarizes, of the 10 studies that I had that did sort of a comparison group design, whether they had a comparison groups, whether they had a comparison group, the number of variables they matched on, and then

which variables.

So, there were two studies that didn't use a comparison group at all. So, that gets more to these before/after issues that I talked about.

Three had a comparison group, but with no matching. So, it was essentially comparing, again, the program schools, for example, with all other schools in the state, even though that might include schools that are much wealthier or much more advantaged than the program schools.

One study didn't really specify. Two studies matched on one variable -- just school poverty level -- and two studies matched on multiple variables, which included school poverty level, previous test scores, and the racial distribution of the students.

I guess I really just want to point out that, even with the sort of limited matching, there were large differences seen for many of the studies.

One study cited a quarter to half a standard deviation difference in baseline test scores before the intervention.

In that case, we are really in a situation where the groups are so different ahead of time, how can we be confident that the differences seen after the intervention are due to the intervention and not due to these preexisting differences, again without just relying on the

model assumptions, which we really don't have any way to diagnose.

I am going to sort of skip the next part. In terms of outcome analyses, I have sort of blown through this idea of choosing schools that look similar. Then the question is, What do we do with those schools?

The answer is, "whatever you would have done on the full sample." I think for the next two days we are going to be hearing more about some of these particular analysis methods of outcome analysis, but it might be things like a difference in means of test scores or gain scores, linear regression, interrupted time series designs, although these again generally do not include a comparison group. So, we need to be very clear about the assumptions being made about the control.

Regression discontinuity design, which I haven't talked about, relies on identifying a cut off such that schools are eligible based on some cut off. So, you compare schools just above and just below the cut off.

I will just mention that the important thing there is that the effect is only identified at the point of the cut off, and that is from this Todd et al paper. I don't have time to get into it.

In general, really, the main point I want to make is that, for all these methods, it is important to compare

similar units, again, to just avoid reliance on the models.

So, I just have a few minutes left, and I will conclude with a few suggestions. For researchers, first, again, clearly define the quantity of interest. Clearly define the comparison or the control condition.

Once that is done, select appropriate comparison schools. The schools should be similar in all variables related to the outcome of interest.

This might be school characteristics, pre-treatment trends and test scores, area demographics. This might involve merging the SSASD with something like schools and staffing or the CCD, which has already been done.

There has been work showing that it is feasible to merge SSASD with schools and staffing. I think that is a really important step to take in order to be able to find these comparable schools, or even to know if they are comparable.

Third, I think it is really important to demonstrate to readers the comparability of the treated and the comparison groups.

One of the frustrations I feel is that often randomized experiments will show a table with the treatment means and the control means, to sort of prove that the randomization works.

I believe that randomization works. What I want

to see is that, for observational studies, I want to see that same table. How similar were these schools ahead of time?

For the covariates that I care about, what was the treatment mean, what was the comparison mean. Those are shown, I think, less often.

In fact, in the studies I looked at using the SSASD, I think maybe three variables were looked at in a couple of studies but, really, more information on that can really help readers determine the quality of the study.

Now, some suggestions for the developers of the SSASD, and I am going to sort of jump on the band wagon so far, which is that, again, standard scores, the sort of more complete information on test scores is important, not just levels of the percent proficient at different levels.

In addition, I think it is helpful for the developers to provide ways either to collect more sort of covariate or background information that they can, or to provide ways to easily merge the SSASD with other data sets, such as schools and staffing, to provide that information easily.

Finally, to conclude, the SSASD clearly can be a rich and useful data base for estimating school level causal effects.

There are two points I really want to make. The

first is that you need to have a comparison group, and that is to avoid situations like Lord's paradox.

Two, it needs to be a good comparison group. It needs to be schools that were similar to the treated schools, the intervention schools, so that we can compare apples and apples and not apples and oranges. Thank you.

[Applause.]

Agenda Item: Designing Gross Productivity

Indicators: A Proposal for Connecting Accountability

Goals, Data and Analysis.

DR. THUM: I am very happy to be here. The first thing I would like to say is please regard the handout that has been given out in my name. There have been a lot of revisions since and, in any case, the paper was a very rough draft. So, work is still being done on it.

Today I will be talking and raising pretty much the same points that were not made as well as I would like in that rough draft.

I will be touching on topics that I think that you will find relevant, points that connect the sort of concerns with using the data base we are all here to examine.

What I want to do, however, is to frame the discussion in a really sneaky way, with something I have been doing in the last couple of years, and that has to do

with how to talk about gains in performance or productivity if you have the right data set.

The idea here, as we have seen, in the kinds of data sets we have assembled in the data base for the group to discuss here, we have a lot of issues.

Now, these are large issues, very, very large issues, but if I were to take you to a scenario where we have, you know, the kind of data sets that would be ideal, and seeing what additional issues there are, I would like to, in a way, amplify the concerns we have with many of the current data sets that we are relying on to make comments about federal programs.

This will come up later. One of the questions in any comparison is really what we are comparing on. So, picking up where Elizabeth [Stuart] had ended, we really like to compare apples with apples, but we really can't compare apples with oranges if we are comparing weight or if we are comparing color. That target is pretty important, and I hope to continue to take advantage of that.

As an overview, let me inform you that the context I have been working on has to do with the present accountability regimes at the state level, the district level and also at the federal level, and how we can properly characterize the progress of students in those

schools.

What I want to begin with is basically a statement of the recent thinking in this area as it relates to the ability of current data sets, even with good data sets of the right kind, to point out places where you can attribute school and teaching effects.

While I rely on the authorities like Steve Raudenbush and Don Rubin and colleagues, to say that much of the current work on value-added models, the conclusions and the effectiveness has to be taken very carefully, if they are cast in terms of the ability to pinpoint where the impact is coming from, whether there is teacher, school or program effects.

So, what I want to say here is that we may have a lot of tools, statistical tools, to work with this, even with the right data, but we do not have all the data that are needed in order to make those conclusions.

We also have -- so, even when we have all the right data, we still have potentially a problem pinpointed what the results are. I have revised the slides. So, I am talking a little bit off sequence.

What I would like to suggest in this talk is that conventional analysis continue to be problematic, even with longitudinal student test score data and so, by implication, more problems can be expected with cross-

sectional data and differences in methods, et cetera, et cetera, et cetera, which characterizes the variety of data bases we have.

The conclusions that I am going to try to lead you to is that descriptive measures of productivity improvement for accounting unit -- teachers or schools -- are still valid, given the accountability data.

Here we are thinking about the ideal data set. I am not trying to make attributions of teacher and school effects. That, I tried to separate that as a separate research project, separating into measurement, use of the data and one that is to get to research for program and teacher effect, school effects.

To me, that requires additional information and that is very sketchy at this point in most places. I have been warned, however, that even if I make this work, and people start seeing productivity in schools, comparisons are going to be made. I am aware of that, but hopefully it will not be this model, given we have tried to separate the objectives now of measuring schools on these outcome variables to characterize the program and productivity, and talking about what is making them effective in one way or the other.

So, we really focus on measurement, and these are the structural relationships, which are the causal effects,

for which we have better and better data.

What I will point to at the outset is, from training and measurement, we really should begin with a lot of talk about standardization.

Even for this limited exercise, I can point to a place where standardization on the kind of evidence that goes into an analysis will help very much. I see a lot of comparisons of evaluations effective to schools that take different cuts of the data, to begin with, and people trying, at the end of that, to compare the analysis and the results.

To me, there is just too much slippage in terms of what you begin with. So, in the end, what differences you see, it makes it much harder to understand.

In the course of this talk, I will indicate -- I will try to show how you can build some productivity indicators that address evaluative hypotheses about growth and change.

One of the points I hope to make is that I am on a mission to try and cut down -- you know, all talk about value-added models, which I don't think exist.

It is a hypothesis about what value is being added that exists. So, one is free to choose statistical methods or various comparison methods, but the eye actually should be on the ball, and that ball has to do with

concerns about the changing world. So, there are only value-added hypotheses, not so much value-added models.

Of those I would talk about, in terms of the model, they would largely relate to large-scale applications of mixed effect models. So, it would subsume many of the sophisticated analyses that you are seeing out there in [Bill] Sander's work and also colleagues at RAND and many other places.

We design also procedures to take care of the inference bit of it, and I am going to rely on showing how Bayesian inference helps communicate the results in a way that I think would be much more understandable.

We will start by defining and measuring value, what we are trying to value. Here are some points that I have made over the last couple of years about what the necessary conditions are.

At the end of it, I would also back off a little bit and say, well, you know, we have all these requirements and they include being explicit about the data going in, and how you parameterize, essentially, the response process that generates the outcomes they are looking at, and that is important, too.

I would suggest that to measure change you should estimate gains. Related to a point I made earlier about how to make the causal inference perspective, we have this

counterfactual framework to work with, and that is well and great, but in that framework it is especially useful, when you have the potential outcomes not observed.

What that means is that you have to do your best either through randomization or some of the procedures, trying to make a good guess of what those things are. That is what it is in a nutshell.

In a gain score, you actually have both. You don't need a potential outcome. I thought, going with the gain score, you have a leg up on the causal claim.

Of course, this is the change in a person. It is not causal in any strong sense that you would use in the sense that a program has caused something to happen. That is a separate kind of question.

If you were to talk about change within the person, I think the gain score has lots of conceptual advantage that goes along with it.

Of course, as I said before, multiple outcomes -- this comes out of the background in psychology, measure construct and multiple operations, et cetera.

That has been fairly well accepted except that, in many of the applications of the value added analyses of school progress and productivity, most people are just exploring that possibility.

We employ standard errors of measurement wherever

possible. The idea here is, yes, we rate test scores, one observation per person, and how flippant that might be.

Well, the test scores, if you remember, consist of a lot of probes about where a person is located on the ability scale. Roughly, that might be it. So, actually it should carry a bit more weight than an N of one.

So, if you are able to characterize its accuracy via standard error, then I think we are making good use of the information provided, the differential information provided from testing.

The metric matters for measuring change. Going back to what Dr. Dunbar said, there are good and bad scales, but I think it would be -- I think what I am proposing here is that when you go with an interval scale that is vertically equating, the equating may be not 100 percent of what you want, but it is still roughly in the ball park of something you can actually use.

For all the other stuff, for all the other issues with aggregation and inference, we should rely on a model for it, so that we can properly propagate the errors from the observation through the model to the result and the conclusions we make.

That is how I think some -- my initial objective of going into this research area was. You know, given my training, how can I take the -- implement the objectives

straight through, from the observations all the way through to the conclusions, and this is a realization of that input and opportunities. Then the other things, also, I will be able to keep that black box open, so other folks can look at what we have done for verification and support of colleagues. They will know whether we are on the right track.

The first big thing we need for a system like this, where the stress is on measurement, is to talk about standardization, and nothing is more important, for what we have to do here than is the evidence base, which I call it, that should go into the computation. So, for example, improvement in data, et cetera.

What I have here is the ideal data base, for those of you who might be thinking along these lines. You have year in school, from year to eight, in the first block on the left; it is going from grade three to eight.

We have these student cohorts. Those are the lines going from bottom left to top right. Suppose a legislature says, with an advisory committee decides that, okay, to make a good case for pronouncing improvement gains and changing directions, for example, given the resources and the time line we have to work with, we should try to get six data points. I know most states are not there yet. Some are, a minority.

I am trying to leave aside what is actually happening in the field, like changes to the test, et cetera, et cetera.

I wanted to try to work with this ideal data base and see what we could do with it. The rest will be just a lot more headache, of course.

For this data block, it will move ahead from year to year. As you will notice in the second panel, the cohort they identified for use in the same accounting machinery has moved to the right by one year. They have excluded the observations for year one.

Now, this is just one way of approaching this. You can have a lot of flexibility on this system. Like I say, we will use as much data as we have.

Then you will have to realize that the information is changing, N is changing. So, the standard errors are changing, unless you have a way of accommodating that and not just visual inspection. You might lose something in terms of how easily you can get to a comparison from year to year.

The idea here is, whatever you choose to do, and whatever variations you might take on this basic idea is that this data base will provide a constant ballast.

We have the same information and so, you know, it will stabilize comparisons from year to year. One of the

problems that I worried quite a bit about is how the performance of systems in schools seems to jump around from year to year a lot.

Hopefully, with a larger data base, some more weight to the estimate, it will be anchored better, and from year to year the pronouncement of trends will not be so radically different.

If I am saying, for year one, this school is progressing at a rate of such and such, for the next year that is not going to change a lot. So, we have an indication of direction, and that tends to be much more comparable from year to year, on account of the fact that we have quite a bit of data to work the rates on.

This is just one example. As I said before, one can take many approaches to this. I work with a multivariate in the sense that we have looked at simultaneously the tests for reading and math jointly.

Basically, this approach allows us to talk about those components individually, at the same time, borrow the overlapping variants to make for more, I believe, stable pronouncements about where the school is headed.

We also make distinctions about cohorts in the data. So, the model can be fairly large. Essentially, each cohort of students, those lines going up in the data block, is basically treated as one of those longitudinal time

series that you have normally. Now this is done jointly.

So, for each cohort, we will have as many series estimated for each subject as possible. So, we have this basic model.

You can elaborate it any way you want, but the idea here, if you remember, is measurement. So, we just want to characterize it -- we don't have an ability to correct change over time.

For the second stage, after we have characterized the school, we will talk about between-school differences to take advantage of the strength of learning from other schools.

You will notice also, those of you who have been thinking about accountability in our data and how you analyze this huge quarter million records at the same time, this approach I am advocating actually doesn't do that.

We will talk about the school. The school has a lot of data, you know, going from grade three to eight, going six years. That is quite a bit of data there.

That actually provides us with a good stable picture of where the school is, assuming that we have 98 percent of the folks responding to the tests.

So, there is no reason -- I think that we can gain a lot more by analyzing all the data simultaneously together.

We do that only when you explore between-school differences, and that comes through in the second stage model where we employ a Bayesian multivariate meta-analysis model.

So, that it takes out that difficulty of having a program run for six days and then having to make revisions, et cetera. This actually can be done at a site by schools, as far as they flow to the system.

Anyway, why the focus on gains? Well, we had thought a lot about this, and there are a lot of other folks who have helped lead us to this inclusion. They include [David] Ragosa, John Willard -- who is somebody I picked up a lot of this from.

Anyway, the idea here should be familiar to you. The important thing is that, for us, it is conceptually congruent, and it has a lot of those not so nice properties of gains, that are not really a true concern in our case because we do not just compute a gain score and stick it into the model. We actually use the observations and estimated gains, parameterize the gains in the context of the model.

So, we should be able to actually get away from thinking about the fact that we are actually using gains in it.

I will not say so much about this, but there will

be some of it in my 2003 paper, and also in the paper that I am putting together. I hope that is good enough for you right now. If you have questions, just talk to me.

In contrast, the major alternative is a residual gains score, which is regression of post-test pre-test. This one has several, to me, disadvantages.

One, when you look at residualized gain scores, what you are doing is you are making a comparison. You are choosing standards that are based on a quite different thing.

In a gains score, I am basing my score on myself, from time one to time two. The residualized gain score, I am basing my gain relative to how the other folks improve their performance. They help me come up with a regression plane from which I determine my gain.

For some other purposes, this also has disadvantages. There are some statistical ones, and they do not generalize so easily to longer time series, et cetera. I am being a little muddled about this at this point, but there are good reasons, given the tools we have, and if we have the right data, we should try the first approach, which is the work with parameterizing gains and trying to explain gains.

Now, if you are in a data situation where you do not have the alternative of actually computing a gain

score, then you are limited in the basis for comparison.

So, that is not a bad thing. You just have different information to work with. So, the best comparison you can make is relative to the group.

When you have longitudinal data, that is a different matter. Then it doesn't seem the right approach to go with a residual gain score.

This is a point I was trying to make earlier with regard to the gains, the conceptual congruence with the true gain that we are really interested in.

I encourage also the idea of supporting causal gains under the Rubin-Hall counterfactual framework. It has the advantage, really, in that the potential outcomes are all observed.

Anyway, the overall strategy, then, is this. To measure schools, what we really want is to answer questions like, on the whole, is the school improving.

Well, that is a very general question, and although policy makers like those questions, they can be better sharpened.

I think there are several questions along those lines that can be obtained using this sort of analysis. First of all, the way we are thinking of doing this is obtain a good fitting, basically measurement in the old sense, of assigning a number according to a rule.

Fit a good measurement model for each school, which is a surface, in this case for math, which I am going to show. They constructed a relevant hypothesis for that school.

You know, for example, you have the outcome variable, the vertical axis. You have grade and year. This is essentially that data grid now, that you are looking at.

Now, turn it horizontally, and what we really want is to know the orientation of this plane. That is a first approximation.

So, this is the predicted value based on all the information in the model we use to come up with it. Of course, this does not have to be flat. If you have enough data points and enough stomach for it, you can come up with various representations of how test scores are best described for that school.

So, coming back to the point I made earlier, I am free to choose any model I want for this, but the goal here is to focus on the hypotheses that interest us, and they happen to be of the variety we have started to call value added hypotheses.

Well, how do you do that? Well, to summarize, you get the best data available, smooth it for irregularity with the most reasonable model, and construct from the signal which we know now is that fitted surface, statistics

that address your hypothesis directly.

So, let me introduce those hypotheses that might be of interest to you. There are three varieties, even though really there are only two genres.

One has to do with how cohorts are growing. There is interest in that. The reason why you don't find an arrow going from here to here or here is that the two time points have to talk about change over time. So, we can leave that. There is no problem. The rest of the cells will still be in the model, helping you estimate the entire surface, but the estimates will not apply to all the cohorts. That doesn't make sense.

So, for those cohorts that sort of have enough data going from bottom left to top right, you can ask questions about, for example, the rate, is the gain from one cohort faster than the other, et cetera. That would be one of the questions that we would be interested in as policy makers.

The second set of hypotheses has to do with within-grade change over time. Second graders are improving in school over time, but what about fifth grade?

Now, for all of these, if you have the right factors in the model, the inference framework supports answering those.

It is very much like the good old analysis of

variance type framework that you are familiar with. You have the data, you have the fitted model. Based on the estimates, you construct hypotheses.

The third species is really, under some interpretation of what a scale is able to support, really a different metric.

We talked about the second type of hypothesis, which is within-grade change over time. Now, this has to do with PACs, our fascination with percent proficient, for example, is one of them.

What this model strategy would do is, we have to fit the surface and we have a relationship between scale scores, percentiles and performance levels. We can report the results in terms of performance levels. So, it is not a separate -- it is not something that requires a separate analysis.

That, of course, goes back to this important thing I am assuming, that I have the right scale. I have the right scale in the sense that it has the properties that support that we talked about change being measured on that scale.

It also has relevant norm information attached to it, and also performance standards being set on that scale. Then, with a scale like this, I always thought until recently that that is the same thing companies are trying

to give us, and successfully.

I have been told that maybe we shouldn't be so optimistic about it, but still, it is probably the one scale that gets us further than the others.

Anyway, here is a list of all the comparisons we can make. Having computed, estimated the slopes, relevant to each of those hypotheses, we can actually make comparisons to standards that are derived from the school.

Have performance of the schools, or standards that are mandated externally, the 100 percent proficient goal could be something that they can do comparison.

You would be making comparisons along the same lines I have done before, but this time entering those standards numerically into the mix.

The result would be you will have nice estimates of your set applying those standards, and good, nice probability statements, if you choose to believe the NTIC to go along with that.

This is a slide I have used to stress that, if you have the right kind of scale for the right properties, the intervals, the criterion-referenced scale scores for the analysis.

Then the other scales that attach to it are probably worked on for reporting purposes. So, we don't have to, at the outset, make a choice, oh, are we analyzing

percentiles or analyzing performance levels. So, if you have this, you know, tie in with the scales, it shows up in all of those, if we choose to.

Here is an example of a comparison with cohorts. On the left, the cohort slope is decreasing over time. We can make the claim that that is decreasing productivity in the school under those terms.

On the right is an example of increasing productivity. The cohorts, the more recent cohorts, have steeper slopes.

Here is an example of a school I have worked with. Obviously, a pattern is never that clear. Then you are able to make comparisons, such as, with this latest cohort, which is the one which is most recent, the latest growth rate is such.

The change in growth rate, this gives some folks the idea of what improvement is. You have growth rate and you have change in the growth rate across the cohort. This is estimated for the school in math and reading.

So, how do you read this? Well, if you look, the first row says, the latest growth rate is such. So, it is about 30.79 points per year for math, and 28.24 per year for math.

You have got the question, what is the rate of change or the growth rate over time? For the first one, for

math, is 3.33. So, it seems like on average, each succeeding year you have increased growth rate of 3.33, the set line scores. The .08 is the standard error.

If you ask for some kind of a confidence statement, well, this is the Bayesian p-value I am providing, but it is almost 100 percent confidence that is indeed the case. For reading, it is 97 percent. So, if you press for a number and also a confidence statement, we should be able to provide it.

There is a lot of interest in talking about and characterizing this scheme in terms of adequate yearly progress [AYP].

Well, AYP is basically a standard. First of all, how do you come up with it? Again, these could be external, it could be based on similar schools, or based on our schools performance.

So, AYP can come in multiple forms and, for various of these hypotheses, they could actually be engineered differently and have very good meaning attached to it, be very meaningful.

In any case, for us, AYP will allow basically -- it must take into account where you start and where you should end up this overall target, between the present time T and the mandated time T to reach proficiency. I am thinking of basically NCLB here.

So, AYP must be defined as the growth rate they are placing on the target, given where you presently are, such as YT, the scale score at the final point in time, et criteria, and the current scale score divided by year. So, you get some sense of the rate.

The little box here is just to remind me, if the opportunity arises, to mention this. You know, we don't have to perform this on all the data. We just need to report it using the right scale.

In any case, these are the predicted grade year means based on that model, with smoothing. Because it is a multi-cord model, it is not as smooth as we would like.

So, we have predictive means, and the questions we ask for Q2 -- that is within the cohort change -- it will be something like this.

This is observed, and we are asking here, is the grade one predictive average increasing, and this is the same question about grade four. So, all of this is based on the model for the data contained in the data block.

The improvement by grade picture looks like this, math and reading separately. So, one through five will be the grade levels, and this is how we are changing over time.

Some of these slides are in the handouts you have. Most of them are. A couple are not, and you have

those in reading. Obviously, they are not all uniform, but you can essentially ask specific questions about each grade mapped.

Now, if you ask a standard based question, then it looks something like this. If you have, for a particular grade, and they have observed for four years now, and the data points are given as such, what you need to do now is fit the best description of that progress, and here you need a regression model.

You can pick anything you think the data will support. So, it is up to you, but by and large, these sorts of data bases are pretty plain. To me, given the data, it drives me toward a simple model like this, at least for a first cut.

What we have then, if we have the targets on the right here at time P, we know what the scale score is. We can compute delta hatch sub-4 based on four time points, and this calls lower.

Basically, it is this bottom edge, the bottom boundary of this wedge, and that is a growth rate that I need, given the best guess of where I am now, over here, to where I need to be.

For those of you who care, there might be other such growth rates taking you to other proficiency levels. What is important is the question of, from where I am

standing now, at time point four, and given where I have to go and what rate I have, am I growing at a rate that will put me there, and that has to do with this dotted line with the arrow in the middle of it.

All we need to do is to be able to characterize the probability, that we don't really need NLCB to consider the upper but, if you want, we can tell by the probability that this growth rate is between the upper and lower, given the data.

In terms of what I call AYP NCLB, it boils down to constructing a reasonable base, a target growth rate, and evaluating for yourself or your school whether you, at this point, given the information you have, likely to make it, and the likelihood is given to you.

The calculations will also provide some sense of classification errors. Here is an example of a school for grade three math and grade three reading.

The green line is a projection. The initial data is, of course, based on this segment of the curve and this is the projected line going all the way into the future.

I know I am not very happy making projections even for one year, but those are the kinds of questions you ask all the time.

So, if you have to ask it, I want to provide you with a frame for thinking about it. If you were to make a

comparison about rates, it is reasonable, but if you want to talk about exactly where you are, and talk about your performance here relative to the observed, those arrows are going to be very, very large, but rates seem to be fine.

In that case, you have all the quantities, and you are able to make probability statements about whether the current rate that is characterizing your school progress is greater than that base level you are trying to get, given the data.

For this school, given what you see here in the picture, it seems that we are growing much faster, at a rate greater than the AYP rate, essentially.

So, you are given a probability statement. Over here, then, for this example, it is a crap shoot, even though it is only 44 percent.

In terms of the percent proficient, all you need to do is translate those predicted values through the cut scores into a proficiency level percentage, and you can track that as well. You can report it as that. You don't have to make any analysis any more.

I am going to stop here. I actually have some thoughts which I hope to be able to make for folks about the inference frame here that will allow us to think a little bit about how to make comparisons across schools, and also across different assessments in terms of what I

call the productivity profiles, which are characterizations of where a school is in terms of one of these -- any of these -- various hypotheses, but we are going to see toward the standard the school is making it, and what probability.

The profiles have both, on the vertical and horizontal axis, have basically percentages on them. So, if you are happy with comparing schools on the effect sizes, I think this will be an improvement. Thank you very much.

[Applause.]

DR. BARNETT: We have about 15 minutes for questions, comments and discussion.

Agenda Item: Discussion.

PARTICIPANT: Some comments that relate to both of the papers, starting off with scale scores, being on the band wagon of saying that scale scores are more useful to most people than just the proficiency levels, I would agree with that.

I would also like to add some cautions related to vertical scales. I think there are a lot of different ways to produce vertical scales, and once a scale score is out there, people tend to just plug it into their complex formula or simple formula and assume that they have something meaningful.

If you think about the concept of how many units of growth is it to go from, say, reading a word to reading

a sentence, and it may be in one scaling method it might be 50 points to grow.

Then you have got kids up near the top end of the scale who are going from identifying the main idea in a paragraph to doing something that is more complex, and that might also be 50 points on the system.

In essence, it is nonsensical to be comparing with points of growth that far apart on the scale, and this relates back to the whole concept of doing these experiments and drawing conclusions is that you need kids in the same place where you start off, to say which of the treatments is doing a better job.

If you are trying to compare students who are starting in different places, or over time when you have a school in different places, you can run into real anomalies in terms of the kinds of conclusions you draw.

If you do a modification to your vertical scale, you could draw a different conclusion. I was just trying to add a caution to this idea of using the vertical scales and doing the analyses in ways so that any peculiarities of the vertical scale are not influencing the conclusions drawn.

DR. BARNETT: Are there any other questions about defining change or modeling change for any of the presenters?

PARTICIPANT: Elizabeth, you talked about the appropriateness, or setting the appropriate comparison groups.

In a lot of the settings that we deal in here, it is not always -- we can't always identify the perfect comparison group because of the way programs are given to schools. I wondered if you could sort of address that.

DR. STUART: Yes, that was one thing that I didn't have time to cover. Are you thinking about sort of these situations where, for example, schools that are not meeting some threshold get some treatment? If you are not making adequate yearly progress, there are some cut offs and, below that cut off, the schools get money. Are you seeing that sort of administration of it?

PARTICIPANT: [Comments not caught by microphone.]

DR. STUART: That is definitely the case in some of the papers that I looked at, where sometimes, because of the program itself, for some reason, you can't find comparable schools.

I think there are sort of two things there. Sometimes if you sort of need to do the comparison, it is better to at least compare the schools that are just above the cut off, rather than all schools.

Actually, one of the references, on the last page

of the handout that you all have, I listed some references, and there is a paper by Dehejia and Wahba, which sort of shows a nice example of how you can run into trouble by comparing these low-performing schools -- in that case it is people getting job trained -- with all other people. So, I guess I would sort of just say, restrict the analysis to schools that are more similar.

The other option, of course -- and one of the papers did this -- is to just say, we can't estimate causal effects here because there are these differences. So, we can just sort of be careful when we talk about them, and sort of not follow up with that.

I guess my third sort of three-pronged answer, there are these regression discontinuity designs which take advantage of that fact.

So, they take advantage of the fact that is a school has a score of 49, they get treated. If they have a score of 51, they don't get treated. So, by comparing the schools at 49 and those at 51, you can estimate the effect of a treatment.

That is sort of what I mentioned in terms of that identifies the effect only at 50, usually, but you can at least do that and get some idea, because maybe that is the relevant place at which to estimate the effect anyway.

PARTICIPANT: I have two questions. One is, Eric

Hanushek has suggested that using aggregate school-level data or even higher can be a problem in terms of assessing certain kinds of relationships.

He is, I guess, especially concerned about resource, and has found that aggregation tends to produce more significant findings with respect to resources than is warranted.

Does anyone see any kind of parallel concern with this data base in terms of looking at other kinds of issues? That is question one.

Question two is the follow up on the first point that was made about vertical scaling. Given that we have school averages -- and we are going to be presumably wanting to subtract grade three in year one from grade four in year two -- what about the concern about vertical scaling here? Is this a serious issue or not?

DR. STUART: I think you would give the same answer that I do. In essence, when you are comparing grade four to grade three, as long as you have, in essence, your control group, the appropriate group, that they are starting off in the same place in grade three, then you are fine.

The issue is, when you start to compare, say, the growth from grade three to grade four to grade four to grade five, and that is the kind of issue that becomes more

complex.

So, if you are comparing the growth of kids who start off very low versus the growth of kids who start off very high, that is where you can draw inappropriate conclusions.

DR. BARNETT: The inappropriateness of those conclusions has to do with our wanting to attribute those changes to something about the construct that is being measured, and the construct naturally changes over a developmental continuum.

I think that was your earlier point about the meaning of a 50-point difference at the low end of the scale versus a 50-point difference at the high end of the scale.

That is very important. The only way you can get around it, or maybe not the only way, but you need a different frame of reference than the scaled score itself in order to get around it.

That 50 points is 50 points in some normative sense as an empirical difference. What it means in terms of the underlying skills is the problematic area of inference, and how much to expect that scaled score to change is also somewhat problematic at one part of a vertical scale from another.

We had several questions and a lot of discussion

in this session that really point to what I would say is the critical importance of not just the Ys in Elizabeth's notation, the outcome measures, but the Xs that she called the covariates. That is just not a variable or two in a complex, multivariate matching framework, for example.

That may be a whole array of variables, information both quantitative and categorical about the sampling units that is important for making causal inferences.

In light of that comment and the questions that we had earlier this morning about the characteristics of the data base and getting some clarity about that, I want to use a little bit of time before lunch to turn the microphone back over to David Goodwin, who has brought a data person with him, to answer some questions that came up earlier.

MR. GOODWIN: We had a bunch of questions in the first session about the PBDMI data base that Stephanie and I tried painfully to answer, and it is clear that we weren't quite as up on it as we might have been.

We actually have someone who does work on it, Barbara Kim over here, to come in and answer some of the questions.

As I recall, the questions were things like, why are we not getting scale scores? Why does the department

have more than one parallel data collection effort underway, how does PBDMI work with the states in terms of collecting data? There may have been a few other things like that.

So, I am going to introduce Barbara Kim, who does work on it. I just want to make one thing clear. She works on it, but she is not responsible for making all the decisions. So, be nice.

MS. KIM: I have a couple of notes based on the questions I was told you had and just thought I would give you that information, and open it up if there is anything additional that you need to know.

Part of this is sort of an education lesson in words. We use the new phrase, EDEN, which is the Education Data Exchange Network.

What EDEN is, is a technology that all 52 state education agencies use to transmit electronically data on their schools, districts and states to the department, and it is also the technology we use to store it in a very large data base.

The other new term is called ED Facts, and it doesn't stand for anything other than ED Facts. It is a logo. It is apparently going to be a registered trademark of the U.S. Department of Education.

That is, everything that allows you access to

that data and that data base, it is the Secretary's intent to have a web portal that not only people within the department use, but people in the public, researchers, whoever, can go to this web portal.

There are ways there to look at the data, take extracts and so forth, but its trade name for anything that gets data out of it, it is called ED Facts.

Our old name was PBDMI or the Performance-Based Data Management Initiative. That has officially ended. It was funding that we received for three years from the Office of Management and Budget to do something about the information management in the education system. That is how we built the EDEN and how we did some initial work with all 52 state educational agencies.

Then, to the data we have in our data base, what we have room for is data on every school, every district, every state, and that is hierarchically.

So, if you call up the school, you can see which district it is in, if you call up the district, you can see which schools belong to that district.

We have data in there about the entity itself, for example, what is its poverty level, is it rural, is it urban, where is it located, is it a charter.

We have data about the staff, what is the FTEs [full-time employment status], are they highly qualified.

We have data about the student. The student data is all at the subgroup level, and it varies with which subgroup the data appears in.

Some numbers, like membership, are strictly based on gender and race. Other things, like assessment data, we have the race data, we have gender, we also have stuff about children with disabilities.

So, depending on what data is, it is all at subgroup level, even if that subgroup consists of one person. That is still in the data base, but it is still not community student-level data.

How much data do we have in there? Right now we have about five states that have all their 2003-2004 data in it.

We are collecting 2004-2005. At this point, for all the states, this is a voluntary system. They can choose to send in data or not. Some states are much more voluntary -- like the processes. Other states are a little bit more difficult children.

The Secretary, however, wants to make this mandatory. She is working with the office of the general counsel at the U.S. Department of Education to make this the official data collection.

Until such time as we are the official data collection, we are the voluntary collection, which means

all the other collections remain as the official mandatory collection.

So, you can't have two official mandatory collections. Within the next two years, we should be the official mandatory.

I sat in through part of the first presentation of this section about specifically the assessment data. What our data base will have is the number of students at all the different -- the number of students in various subgroups for each state, district and school at state-established proficiency levels on state assessments.

I understood for your purposes scales would be better. In this case we have proficiency levels, because that is what the department needs in terms of managing No Child Left Behind.

The legislation is about proficiency levels, although our proficiency levels are sometimes more than just the three, above proficient, proficient, and below proficient. It is highly unlikely that it would be changed through the scores.

The other part that was brought up in that presentation was other variables, and that is something that we will have.

When you call up a school, you can find out, and for the five states that do have all their data in, you can

see what is that school's poverty level, is it a rural school, how many teachers did they have, how many of them did they classify as highly qualified.

You could call up and see their membership by the race and gender. You could see how many children with disabilities they have, what type of disabilities did those children have.

So, all the other factors that you would be looking for to find comparable groups would be in there, but the bit about the scores would probably not be.

The bigger part about what we are doing is that it is a very ongoing process. When you go out into the states, there are some states that are very sophisticated in this area, where for years, like Florida, has had its new information system. They gather everything electronically, and they have well-trained people in the schools.

There are other states that engage in a lot of paper transactions and counts of this. There are several states where the data provider person is not treated with any respect. They have the office in the basement.

One particular state is attempting to get an association of data providers, so that they have some clout in getting the training that they need to do the work that they do.

Over the next years, our approach to this is to work with all 52 states, the 50 states, District of Columbia, and Puerto Rico, so that the people in the schools that collect this data are professionals, that they have the technology they need to gather the information, transmit it to the state education agencies, who will then transmit it to us, and that we would have the technology so that you could follow up on our web page and get the data that you need. Hopefully, that answers a lot of your questions, but are there other questions that you have for me?

PARTICIPANT: [Question not caught by microphone.]

MS. KIM: The question is about what information we are going to have on the assessment data that is in our data base.

We use a term called meta-data. What meta-data is, is an explanation of the numbers in our data base. So, for state assessments, there would be a place where you could call up.

For a particular state it would say, this is the assessment that was used. There would be some other information about that assessment.

Whether it was down to all the detail you would need, it may not, but it would, at least for a state say,

the scores for the third grade were from this particular assessment instrument. So, it would give you information on which to go find the rest that you would need.

PARTICIPANT: [Question not caught by microphone.]

MS. KIM: What we would ask for from the state is the background on their assessment data, and that would probably be one of the items.

One of the things we are going to be doing hopefully more and more is working with PPSS in terms of the reporting out and some of that other type of information that we need, so that the data that is in the data base can be used for different purposes. Is that a closer answer to your question? Are there any other questions? Okay, thank you.

DR. BARNETT: Thanks very much. As the moderator, I would just like to set the stage for later today and hopefully some discussion tomorrow.

I would hope -- I think of this in terms of policy-maker questions. I work a lot in early childhood. I know one of the policy-maker questions that I know gets asked all the time is, states have different preschool education policies. What are the consequences of those for third or fourth or fifth grade test scores?

In view of the concerns about aggregation bias on

the one hand, and concerns about -- they are not interested in program impacts. They are interested in where differences in policies across the 50 states, which is a state-level difference, has state-level differences in test scores.

What do you make of that? Are policy makers asking for something that is just not realistic to answer, or is there some possibility with data sets like these, and what would be the requirements?

Yes, they may be defined in a certain way right now, but if policy makers then could get information that says basically either this will never answer your questions, or your questions could be answered, but your data system is going to have to have the following characteristics, I think that kind of information could usefully modify what is happening with data collection. So, thanks very much, and we will take a break for lunch.

[Whereupon, the meeting was recessed, to reconvene following lunch, that same day.]

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

DR. BARNETT: We will ask everyone to introduce themselves, and we will start right off with Michael Scriven.

Agenda Item: Can We Infer Causation from Cross Sectional Data?

DR. SCRIVEN: I am Michael Scriven of the Evaluation Center at Western Michigan. I am going to talk about one issue from the foundations of this research area, the issue of experimental designs to establish causation.

There are others that deserve attention. For example, there is the problem of distinguishing real effects in the data from the artifacts of high-stakes testing, which I think hasn't received enough attention yet.

For example, one of the artifacts of high-stakes testing is the redefinition of proficiency, and the redesign of the test content and the item distribution across topics and, of course, the effects of administrative fraud, especially in controlling the population of test takers, by arranging that the poor performers stay home on that day, and by correcting student answers.

I am not sure how familiar all of you are with the extent of that. The evidence from the two NSF studies is that it is stunningly widespread.

That is, it is occurring in far more than half of the schools that are taking the tests, and it is on a fairly large scale. That is, teachers are just changing the answer sheets substantially.

That effect, as it currently looks, may alone out-perform all results of NCLB. So, I don't think it is sensible for us to be proceeding too far with the technical analysis until we have pinned down the reality of the effects as effects of the interventions rather than as effects of the effect of the political reality of the intervention.

I am going to talk instead about the general issue of causation. I got a nice lead in from Elizabeth Stuart, who started by talking about the randomized experimental control trials [RCTs] process as the gold standard for causal investigations.

It is not the gold standard. It is, at the very best, the gold plated standard, and even that makes me feel a little nervous. I think it is the eight caret gold standard, almost entirely dross.

As Tom Cook said in print lately, in the educational field -- and the same would be true in most of the sort of social service investigations fields, it is completely inappropriate to refer to RCTs as the gold standard because of the problems of treatment transfer, the

leakage problem, the problem of differential attrition, the political pressure problems, the ethical barriers that are legally required to be observed, the cost problems, the time window restrictions, and so on.

In my view, there is another reason for avoiding the term gold standard in this connection, which is the double Achilles heel of the design in the educational of social field.

The fact is, as it is normally practice, a zero blind design. It is not even single blind. It is zero blind and, hence, susceptible to the counter-explanation of the Hawthorne effect and its evil twin, the counter-Hawthorne effect, the control group competitiveness phenomenon.

Now, that is pretty serious because people constantly talk as if somehow the use of the RCT design in pharmaceutical research, where it is indeed the gold standard, is somehow the same thing as it is used in the educational and social services research field.

It is not in the least the same because, in the pharmaceuticals field, it is always meticulously double blind.

There are also other major differences between the two which can be expanded on at some length, but the general point is, in our field, it is not double blind, it

is not single blind. It is a pale imitation of what is, in another field, a gold standard. Therefore, we should be extremely careful about talking as if it was more than that.

Now, I am going to go back now to my paper, as you have copies of it, which is entitled, "Can We Infer Causation from Cross Sectional Data."

One answer to the question in the title is, certainly not, if the only scientific basis for causal claims is the results of experiments, in particular, experiments with random allocation and treatment between the subjects and the control and the experimental groups.

Equally certainly, however, that kind of inference is made all the time from science, and inference from cross sectional data to cause conclusions.

For example, from the cross sectional views provided by magnetic resonance scans, to conclusions about brain tumors, and by the cross sections cut by the scalpel and saws of the forensic pathology performing an autopsy to the cause of death.

Indeed, there are simpler examples. Looking at the cross section of an ancient water cypress, one can infer, from the rings and shape and thickness of the rings, to the occurrence of a big drought in a particular year, which almost sucked water out of the swamps.

My task here is to look with some care at such examples and see if we can learn something from them about what can and can't be done with the school based data base, either in its present form, or in some revised form.

Now, I am going to skip heavily through this, because I am anxious to make sure there is plenty of time for questions.

I go through the autopsy example in some care, to show how, in fact, the process of eliminating alternative causes of death is done by looking at the details of the way in which potential injury -- an injury that was potentially fatal -- occurs.

It is this looking at the causal chain that enables one to eliminate other possible causes, and finally pin it down as being due to a gunshot.

So, that process of using the modus operandi evidence, the footprint evidence, to eliminate other possible causes, is characteristic of the use of this particular kind of approach.

This approach is, indeed, absolutely universal and common sense, and throughout the disciplines. In particular, it is worth remembering that it is the standard approach in the oldest discipline that we have in any organized form, which is history, which is now more than 2,000 years old, when people began to talk seriously about

the proper methodology for identifying causes in history.

The recent publication of a four-volume collection of the historical papers in this area reminds us that there has been a very extensive discussion of this for a very long time. We now are in a position to talk in some detail about how this type of causation is established.

Well, of course, it is never established using control groups or, for that matter, not usually using comparison groups, except ones that are very artificially constructed.

So, it is clear something is going on there. It is either charlatanry, and they really aren't entitled to any causal claims at all or it is, in fact, the use of another method for establishing causal conclusions.

It seems to me, given the huge range of alternative sciences, in which non-experimental evidence is used to establish causal conclusions -- for example, epidemiology, astronomy and large aspects of biology and so on -- one must inevitably conclude that there is no preference at all for the idea of RCTs for this.

It is clear that the reason that was given by the earlier speaker for this, which is the reason most commonly quoted, that the RCT, because of randomization, eliminates the possibility of all other causes, is clearly not true because, of course, the work on effects is present. So,

that is one case it is not excluded.

In general, the rest of that recipe, which is that once you try to eliminate all possible causes, is of course perfectly correct, and that is exactly what you should do, but you don't need to do randomized control trials to do it.

You can do it as the historian does or the epidemiologist does, or the forensic pathologist does, by getting down to work, looking at possible causes from the first evidence that you see, and then tracking back through that list of possible causes, looking at what traces they would leave if they had operated, and seeing whether those traces are present or not, until you have reduced it to one or two or three, in which case you might have co-causation or similar cause.

Now, can we transfer that over to the evidential data base that we are talking about in the schools situation?

Clearly we can, and I talk about special cases where you can, where some short term traumatic event occurs, and you see a big blip in the scores or a big drop in the scores, and of course we can extrapolate back to the slower acting causes, provided that we can actually establish the reality of the underlying change, which we are assuming we need to explain.

So, I don't think there is any difficulty here in principle. If you look, for example, at the very good RAND study that was recently released, done for Carnegie, in which they evaluated the value-added model as an approach to teacher evaluation, an arena in which it originated, and Bill Sanders in Tennessee started it off, and Dallas picked it up with Bill Webster's application of it in a much more sophisticated version.

It is clear that in that context their conclusion is that this is an interesting approach but there are still four or five alternative possible explanations that need to be definitively ruled out by further study, and that is exactly the procedure that we have to follow in our field.

We have to look at whether or not the particular alternative explanations -- Bill McLaughlin mentions some in his paper -- can be ruled out. Basically, they can often enough, with some work on the ground.

Now, we can't turn loose some huge army of case study researchers in each state to do this for every school in the state.

We have to do what we have done on many other occasions, which is pick and choose our schools with some care but then, having done that, we can get some basis for drawing a tentative conclusion that we are getting some effect from an intervention, if indeed it does turn out

that we can eliminate the other possible causes in these schools, where you would, of course, pick them from relatively stable populations and relatively comparable SES and so on.

Then, if you get that going, then you have established that this particular intervention is capable of producing effects in a relatively simple sample of schools. Then you can begin to treat it as one of the possible causes when you start looking at more complicated cases in the urban schools, where it is always trickier. Then you can still proceed in the same way, and you will in many cases be able to do something.

This is a generalization of a success case method generalized to cover the case of failure as well as success, and to look not just at what he does, training interventions but also other forms of educational interventions.

So, we have quite a lot of literature on this now, and I think it is clear that that is a perfectly feasible approach.

The other feasible approaches were advocated by various authors of the papers that you have with you, including, of course, interrupted time series, which can be stepped up to the point where you use random time intervals for the interrupts, and random duration for the interrupts.

So, you can get a result out of an interrupted time series, which it is really absurd to suggest could be due to anything other than the interrupt.

If you randomize for the time of interrupt and duration of interrupt and you are getting the results -- for example, suppose you are testing a particular method of reinforcement for learning a particular skill -- then you can do this quite well and get a nice example of an interrupted time series designed to demonstrate that, in fact, when you do switch to using visual task auditory reinforcement you get a substantial gain in their learning speed.

Now, the same can be said, regression discontinuity, the one quasi-experimental design which is allowed by the RCT people as a possible alternative in cases where you can't do full RCT, it is clear that there, too, if you are lucky enough to be able to meet the statistical requirements, which are very severe, and if you have the skills to do the work, then you can also produce a result which is really inexplicable in any way except for attributing causal power to the intervention which has led to the change in the regression rates.

So, we have got a number of straightforward possibilities. It is worth remembering now it is tempting, when you are looking at experimental design, especially if

you sort of draw it up on Cook and Campbell, to look at lines of Os and Xs and think in these terms, and then think how nice it is in the RCT design to have eliminated all possible causes.

Of course, if only the thing would sit still and look like Os and Xs, but unfortunately, of course, there is this inconvenient time consideration.

That means, thinking back to the Title XX evaluation which I was head of the meta-evaluation team for, in which perfectly designed RCT design proceeded to eat up \$8 million without the slightest trace of a result because of straight leakage, in this case manipulated leakage, that is, the superintendent in each school district in which they had agreed to have a control school and several experimental schools, the treatment basically being that you had to use one, four or five optional interventions that would be funded by the feds.

What the superintendent of schools, at the meeting where he convened the principals, he said, okay, guys, we have agreed to play along with the feds on this one.

So, you five principals are going to get a nice little pot of money that you can use on any of these five interventions, and you, Jack, are going to be the unlucky control group school, but don't worry, I have got other

sources of money.

Of course, it took about three months before we completely balanced the control and experimental group, and there was no trace of any results at all.

Well, that was a case where the feds had designed the complete experiment. It wasn't the case where the person who got the contract, which was STC, had made the mistake of designing a rather easily-breached RCT design. It was a case where the feds themselves were very smart -- a federal officer, in fact, had designed this and got caught on it, and blew a lot of money and time on it.

That is worth remembering. We have a lot of experience, and I talked to Barry McGore about this the other day. He is now the education director for the European Union, which runs an enormous number of causal investigations.

He said, yes, well, I mean, I remember very well what happened in the late 1960s and 1970s and 1980s when we were really hog wild for RCTs when we could get them, made a real effort in them.

Maybe one in 10 times we got the semblance of a result, and nine out of 10 times it went to hell on us, either through differential attrition or leakage, leakage being the big one, and leaking including, of course, manipulation by the people who had the strings on the money

and so on, and of course also including all the cases where teachers get together socially or kids get together socially and talk about the exciting things they are doing and they tell their teacher about them, and the teacher tries them out and there goes the difference.

So, there are problems, then, that we need to be very aware of there, and there are many situations in which the alternatives to RCT are greatly superior, not just second best.

The term quasi-experimental, like the term proficiency, was a sales job in itself. It sold you immediately on the idea that it isn't quite the real thing, but it is sort of halfway there.

Well, it is not quite halfway there. It is a long, long way from being there, and it is not the only way of getting it.

I conclude the paper by shocking probably everybody by saying, we have left out two. We have listed five good designs that you can use for causal investigations at this point in interrupted time series, regression discontinuity, three versions of RCT and so on, but I add two more.

The one that people understand pretty clearly is the theory-based one where whole disciplines like astronomy, which have never run a single experiment with

any star in the sky, are quite happily able to produce causal accounts of the genesis of black holes or pulsars and so on, and do so with very good reasons which are, of course, theoretical reasons based on a theoretical analysis of the phenomena they can see in cross section up there.

The interesting thing about astronomy is that, except for the astronomy of the solar system, all observations across some trivial period like 2,000 years are cross sectional because things move rather slowly, or they move rather fast, but they don't produce much relative motion.

So, we have plenty of sciences like that, the greater part of geology, and other parts of sciences like epidemiology, where we have plenty of alternative approaches, and which are very heavily theory-based.

My favorite example is, of course, the smoking and lung cancer case, where we never had the experiments, but we had absolutely convincing evidence, rightly convincing evidence that, in fact, smoking was a cause of lung cancer, and it was a dosage-based natural experiment that were analyzed by clever analysts that produced that. It was not anything like an RCT. So, we have plenty of that.

Then I add the objectionable final remark, which I was greatly encouraged about, by finding two lines in

Cook and Campbell where I am sure it was, in fact, Don Campbell, who wrote it, and I am not sure Tom read it when it finally went to press, because they look very much as if they were added after an evening of confrontation with his conscience.

He says: "Still, at the end of the day, I have to concede there are times when people see causes operating, and rightly conclude that they are seeing something cause something else based on the evidence of their eyes."

This is, of course, the way that all of you use the term "cause" in your ordinary life. It never occurs to you for a minute that there is some doubt about the fact that, when you put on the break pedal in a well-maintained car, the car is going to slow down.

You are quite right to be confident that that is true. Once in a while you will be wrong, but that is the nature of science. It is not the nature of some flawed design. It is observational, that you know very well that is what it is caused by.

Now, I am going to conclude by reading to you a short extract from a paper that you won't have seen, which is written by Eleanor Chaminsky, whom many of you know ran the balancing act in the General Accounting Office for a long time, which was the office of program evaluation, et

cetera, and managed to achieve a degree of credibility in both sides of congress that was almost unmatched, in doing the evaluation work that she did.

She did it because she was an extremely savvy person, both politically and intellectually, and in terms of scientific training.

Here is what she says at the end of this paper which will come out -- it was written in August, and it will be out in an issue of New Directions in Evaluation, probably in a month or two.

Here is what she says, having listed all the drawbacks about the experimental approach, the RCT approach, and is complaining about the fact that the education department has indicated preference for this:

"Given so many negatives, both to the education department and to the field of evaluation, the department's priority to experimental design is not readily understandable."

She concludes, finally, in the last paragraph of the paper, as follows: "One can hope that this priority represents nothing more than an agency enthusiasm for a particular method."

Even so, enthusiasm has no place here. Randomized studies have always been careful, well-reasoned, scientific approaches to specific questions.

This is no one-size-fits-all methodology. Evaluators need intellectual freedom to do their work, and their freedom involves the continuous consideration of alternative thinking, alternative choices, alternative solutions.

In the end, the paradox of the Department of Education's priority to the experimental design is that it is non-experimental. It reflects the introduction of ideology into the heart of the empirical paradigm. Thank you.

[Applause.]

DR. BARNETT: We have decided to shorten up the talks just a hair and have 10 minutes of questions after each speaker, rather than holding it all to the end.

PARTICIPANT: [Comment not caught by microphone.]

DR. SCRIVEN: I haven't written up a long one. I refer here to the list that is provided in the RAND study of experimental design, and I haven't got it with me.

PARTICIPANT: [Comment not caught by microphone.]

DR. SCRIVEN: What is interesting about all of this, in a way, is that it is clear from what Tom Cook says, in his paper on answering objections to the use of the randomly controlled trial, in which he rightly refutes a lot of the really bad arguments that have been given against it, he nevertheless does a very good job of listing

the good arguments there are, that have to be met in a particular case.

It is clear that one can do this on the ground, but the skills to do it are the same skills that you need in running an RCT to guard against the counter-explanations, the attrition, the leakage, the interventions from outside, political pressure and so on, which Eleanor outlines at great length in her paper, by the way, if you want to see an essay on the political influence on RCT design.

She pulls up case after case from her own experiences at GAO, where there was direct intervention, direct RCT designs. Because they are around for quite a while, they are very vulnerable to this.

Now, the situation that Cook is calling attention to could be summed up this way: No really good quantitative researcher, interested in using RCT, can do it without really good qualitative research skills.

Either they or a member of their team has got to do this subtle transition effect, and catch it before it gets out of control.

Leakage can begin and be stopped, or you can drop off that particular school from a study, but you have got to move pretty fast on it, or you finish up looking at the outcomes before you decide who to drop. As was well put

this morning, that is not exactly the way to do it.

I think that one can clearly do a very good job on this with care, but you need skills, and the skills are basically the standard qualitative research skills, of detecting by talking to people and using structured interview techniques and using focus groups when the beginning of trouble is beginning to emerge, and that is sort of part of the tricks that are involved in doing good RCT, but they are equally as good for the general elimination method which I am describing as the main alternative.

So, I was asked the two general points. Number one is, it absolutely won't do to suggest that the way we use RCTs in the educational area is a gold standard. As it is normally done, it is zero blind.

So, it is not a good example. Also, by the way, I think you can do double blind studies with RCTs. It requires the same kind of ingenuity that the RCT supporters rightfully complained about as lacking when we went through this period of 15 or 20 years when a whole lot of studies should have been done using RCT, which were done using really sloppy qualitative methods.

So, I am entirely on the side of the RCT enthusiasts about that mistake, but the other mistake can be made in the other direction and I think has been.

So, one has to be very careful of separating out these issues. So, one of the issues is not to get over enthusiastic about RCT. The other is to be open-minded about the way in which other methods can produce conclusions beyond reasonable doubt.

The interest here is not in being theoretically bullet proof, but in being practically good enough to bet your life on. The standard here is the standard used in felony lawsuits, where it has got to be beyond reasonable doubts.

In non-felonious misdemeanor lawsuits, it is balance of evidence. That is too weak for us. We are after beyond reasonable doubt, and that is quite right.

You can get to beyond reasonable doubt, as every law court sees many times a week, by all sorts of methods, to establish causation among other things, by using methods which I have been describing as the general elimination methodology.

PARTICIPANT: In terms of the SSASD, do you feel that the right kinds of information are in the data base to be able to eliminate the alternative hypotheses?

DR. SCRIVEN: No, as I say in the paper, for the most part, the convincing elimination of alternatives will require some work on the ground.

That is why, to be realistic, since we can't do

that throughout the whole state over the 99,000 schools that we have got in the data base, we have got to pick our sample.

Now, that is something we have been doing a lot in some of these studies anyhow. So, we pick a sample. Maybe it is 20 schools or 40 schools, and we put in a group two or three people for two weeks or so, on two or three occasions, so they can pick up earlier or later, new causes that could be emerging and competitive.

Then you have got a combination of the existing data base plus some extra data. That, I think, is typical. Now, there are cases, obviously, the traumatic cases of the Gulf Coast hurricane and the Columbine school problem, where you can do it straight off the cross sectional data. It is not going to be hard.

Therefore, it is plausible to say, within cases where you have got very large swings, I think larger than Iowa in the middle there, you might well be able to pull that off from the SSASD data base, but you can't in general.

DR. BARNETT: Thank you.

Agenda Item: Controlling for Student and School Differences: Value-Added and Residual Gains Approaches.

DR. MIRON: My name is Gary Miron. I am chief of staff at the Evaluation Center. While Michael [Scriven] is

very philosophical, I tend to be very practical. While Michael has been very busy writing books about evaluation, I have been busy grinding out technical reports that pay for the work that we do developing a theory and method of evaluation.

I thought I would draw on some of the technical reports that I have been working on over the last eight years I have been at the evaluation center, and many of them have focused on looking at school choice and charter schools.

So, I was going to draw upon those. When I got involved in some of the studies at the evaluation center, I was really quite surprised to look at other studies, my other peers who were also researching the topic, and found such a wide variety of studies.

They varied not only in the quality of the designs, but in the sensitivity of the measures that were used.

They varied in scope. They varied in the number of years that they covered. They also varied very much in terms of the controls that were used, the number and nature of the controls that were used in the studies.

So, I was really quite surprised to see such a variety of studies. Charter schools, we have had charter schools for some time now. The first one opened in 1991,

and most of the large charter school states now have more than a 10-year record. Yet we still don't have a definitive answer on whether charter schools are working.

Anyway, I wanted to talk about some of my struggles in working with less than desirable, or I might say, best available data, because that is often what we are required to do when we are contracted by state agencies to come in and conduct evaluations.

I am embarrassed to say that I have been involved in doing some of the studies with weaker designs, and I am also happy to say that I have been involved in a few studies where we have been able to match students across the state and track changes over that time, for that student doing a quasi-experimental design.

So, I have been involved in a number of different studies, and interestingly, we have identified a few of the more weak charter school studies, or charter school reforms, and we have gotten criticisms for our work.

On the other hand, a few of our studies have been among the most positive in favor of charter schools. So, they are all over the board, and certainly they were all over the board in terms of the quality of the design and so forth.

When I looked at the group of studies out there, including our own, I kind of grouped them into a number of

sections here.

Some of the weakest studies have been the cross-sectional studies, and then there have been those that probably the most common have been looking at successive cohorts with and without controls over time.

Then, same cohorts. Then, again, those designs one to five, that grouping from one to five, those are working with individual student data.

As many of you know, in many of the states, even states that have individual student data, we haven't been able to get our hands on that data.

First, let's look at some of these cross-sectional studies that have occurred. A couple of noteworthy ones were released last autumn, the autumn of 2004.

One was by the AFT [American Federation of Teachers], and they were working with NAEP data, and they did a cross-sectional snapshot of charter schools relative to non-charter schools.

They did some blocking of the data by background information, such as percent reduced lunch, ethnicity and urbanicity as well, I believe.

That was an interesting study, and they found that charter schools on the whole weren't gaining -- weren't performing -- as well as the matched comparison

group from the traditional or non-charter public schools.

A couple of days later, Caroline Hoxby released her own study, where she claimed to have captured 99 percent of the charter schools across the nation, also a snapshot, cross-sectional picture in time.

Not surprising, for some of you that may know her work, it was very positive in favor of charter schools, again, looking at one particular grade level at a time, and not looking at changes. It was, again, just a snapshot.

These are two relatively weak studies in terms of design, but they garnered very much media attention, which is pretty exciting, pretty surprising.

As an observer, it was very fascinating because there were, of course, a number of other studies. I think what was unique about those was that they were national studies, that they did look across states, and we haven't had those cross-state pictures yet.

There have been other -- typically we see media outlets, and often advocacy groups are often involved in doing these cross-sectional snapshot reports or designs.

They do those because they are cost effective and easy to do. A few state departments of education have also done these cross-sectional studies as a way they evaluate their charter schools.

If you go a level up to the successive cohorts, I

think this is the most typical group of studies I have seen on charter schools.

That is basically looking at successive groups of students at the same grade level, fourth grade readers in 2000 compared to fourth grade readers in 2001, and looking at those changes over time and with a time series. Some of those studies, of course, have controls, and some of them don't have controls.

In that group of studies, we have done a couple of things. One was a residualized gains analysis that we did in Pennsylvania, a state with 77 charter schools at the time.

We had scale scores for the state. What we did is, we realized that the demographics within the charter schools, because they have high levels of mobility, that the demographics in the schools are changing. So, there are relative comparison groups that we would want to have that change from year to year as well.

So, using a residual gains approach allowed us to be sensitive to the changing comparison or changing predictive values that we would have over time.

So, on the whole, charter schools in Pennsylvania tended to attract lower-performing students. So, the residual scores, on the whole, were negative scores over time.

So, with that, one could look at the data and conclude, hah, the charter schools are doing poorly. In fact, the charter schools in that state tended to attract students who were performing less well, even after controlling for demographic and geographical factors.

Now, when we looked at the residual gain scores, they were actually gaining slightly against the traditional public schools, which means they still weren't performing as well in absolute scores, but over time they were approaching the regression line. So, they were doing better over time.

Another analysis that we used for an Edison study that we did, we had two types of data that we used. One, we had individual student data that we could do longitudinal analysis on, but we also wanted to make cross-state analyses.

So, we used something that is called an odds ratio analysis, and I discuss it in my paper in greater detail. It is a methodology or strategy that is used often in epidemiology.

It basically looks at the odds of passing or failing over time relative to a comparison group. That gave us a way to use cut scores and have a similar measure across states to look at changes over time.

If we look at same cohorts, at six and seven, in

our Connecticut study, we had a four-year study in Connecticut.

In that state, they were very, very proud to let us know that they had individual student data, and they could track any student no matter where they went within the state.

So, we were really excited. We were going to finally get our hands on some individual student data, and yahoo, that was going to be fun after mucking around in less than desirable data for some years.

As it turned out, the regulations and restrictions wouldn't allow us to have access to that type of data. So, what we did do, though, we looked at the same groups over time, and this is a study that I have seen a number of researchers use, which is to compare with a comparison group, but to look at the treatment group or the charter school group, fourth grade readers in 2000, sixth grade readers in 2006.

So, we could basically look at that same group and try to limit some of the mobility over time but, again, we couldn't eliminate all of the alternative explanations for changes in those results over time, but we could see over time -- in Connecticut we used both the successive cohorts and the same cohorts and tracked them over time and found that, in fact, the charter schools were making larger

gains than the comparison districts and comparison schools.

If you go up a little bit higher, those study types from groups one to five all work with individual student data.

If you look at number two, we have been doing -- we are in year two on a statewide study in Delaware. Delaware is a very cute state. It is very nice and it is very small.

They gave us the whole state data. I mean, we had to go through clearances, of course, but the whole state, seven years of test data for every student in the state. This was exciting.

We got to do a matched student design, which basically meant that we created those strata based on student characteristics.

We grouped them by gender, ethnicity, Title I status, special ed status, limited English proficiency. We blocked those up.

Then we matched them, and we picked from all the students in the state that had the exact same characteristics. We randomly selected a matching student, and then we compared change over time.

Actually, we looked back in time. Students, to get into one of the six panels that we created, had to have a test score in two events.

We wanted to get three events, but because they were not testing and did not have that type of test data available in every year, the three test events would have been too long, and we had too few students that we could include.

We did get six panels designed and got to track students, and we covaried by previous test scores. So, we could even control for relative performance levels.

In that study we found that, on the whole, charter schools were gaining more than traditional public schools, but in that state there were extremely large differences between the schools, with some performing quite well and others performing less well.

What was also exciting about this methodology, there was one particular school that is very homogeneous. At times it has been reported that it is using an admission test to get in and only the top scoring students are getting in.

I saw that and I thought, wow, with our controls, when we get these in place, we are going to be able to explain the phenomenal impact that charter schools are having on these students because we can control the data.

To my surprise, after -- we could see that clearly the school was taking in very high-performing students. They were performing very well on the test

results before coming to that school. So, it was true that they were attracting or selecting the top students.

What was fascinating is that that methodology allowed us to see that they were still making gains larger than what would have been predicted over time. So, that was interesting and probably can be explained by that very homogeneous focused learning community that was being created.

That was one example of a study that we could work with individual student data. Randomized experiments or randomized field trials, we have made one attempt, and that was in our Illinois evaluation charter schools, where we tried to get the waiting list from schools and try to draw students from that group, because we would have had some sample of students who would have gotten into the school but, because of the lottery system, didn't get in.

Unfortunately, we couldn't pursue this because a lot of the schools didn't audit the waiting list, and often the schools were reporting waiting lists of 100 to 200 students when, in reality, there were only 30 or 40. Because of that, we weren't able to pursue that methodology.

Carolyn Hoxby did. She did that with a subgroup of the schools in Chicago. That is the one study on record that I know on charter schools that is using a randomized

field trial.

Aside from that, Mathematica is doing a study for the feds which is, I think, going to be one of the most promising studies, but I don't think it is going to be a definitive study.

It is a very promising study where they are looking at charter schools in, I think, around 50 schools across five or six states, and they are looking at between two and a half and three years, looking at the impact of those charter schools on students who are randomly assigned to public school or a charter school. So, that will be interesting when those results come out.

So, some of the limitations in the data that I see -- and in most of my studies, I am using the state data, but it is very much like the data that we would see in the SSASD.

Some of the limitations that I often come across working with charter schools, which are smaller schools, is that with the exclusion of small groups sometimes, where the states don't report test data for small subgroups.

Something that has really surprised me, too, as I have been getting into it is the norms that are changing over time with state testing, the tests that are actually changing themselves.

Then, of course, a problem we face when testing

these schools is the high mobility rates. So, when we are working with group-level data, it is very hard to control when you have schools with high mobility.

Some of the fixes or improvements, just by adding some of the fixes, just by adding more groups that we can test over time, the more grades that are being tested means that we are going to be able to do more rigorous designs, which is pretty nice.

That is happening as a result of NCLB, where states are forced to implement more testing at more grades. I don't want to say the scaled score again, but it is so important to have scaled scores relative to these cut scores, which are very crude measures, and don't help us capture the sensitivity or the movement that happens in these lower performing cut offs or subgroups, like whether they are basic or advanced or approaching or whatever the states call those categories.

There is movement there. I know that Connecticut did have an index that they created which was based on the cut score, the percent meeting state standards, and then it included the percentages that were shifting from the lower level, from basic to approaching. Scaled scores really help us look at change over time.

Then, I mean, this issue of what we can do with individual student data, and what we did in Delaware was

just phenomenal.

For the amount of effort that didn't vary from what we did in other states, I think we were able to provide a better answer, a more certain answer, and control more of those alternative explanations.

Jumping to some policy issues, what about cost for these studies? I have just highlighted some of our studies. So, our studies, if we look at just the test analysis component of our state charter school study, in Michigan, you know, we had a contract for like \$150,000 over a three-year period, but only \$10,000 was devoted to analysis of student achievement data.

This very rigorous design that we could do in Delaware cost about \$12,000 in year one for the time and labor for that component of the work, which is very inexpensive for such a rigorous design, where we can track students back. It is fantastic.

Most of our studies, for that component of the study, it is typically \$12,000 to \$15,000 one can do that type of analysis.

Now contrast that with a Mathematica study which is, I think, over \$5 million. I think in that year the feds had close to \$6 million devoted to charter school research.

When we devote such a large chunk of our work -- granted, I think the study, when it comes out, because of

its design -- gold standard, gold plated, whatever -- it is going to be one of the best studies that we have, but I am not certain it is going to provide us with a definitive answer yet.

Coming back to that definitive answer and raising the question, what is the appropriate way to evaluate the impact of federal programs? Before coming to the evaluation center, I spent 10 years in Sweden.

In 1992, I got involved in evaluating the national voucher reform. It was so fascinating there. First of all, from a policy maker's perspective, it was a very centralized system.

They flipped the switch and there was national reform overnight. Contrast that with what we have here, where we are trying to evaluate charter schools and get an answer, and yet every state has a very different implementation.

Switch back to Sweden. When we did that evaluation, we had a national standard, a national curriculum, national tests. It was pretty fantastic. Believe me, we still had lots of issues to fight about, about testing, about whether weighting the testing was being done appropriately from year to year. Also, we had the struggle with the government to release individual data from school level data, which was the unit of analysis for

our study. It was fantastic that we had that uniform data across the data. Again, come back here and every state study that we do, every study of charter schools has a very different and unique set of limitations.

So, the policies, the way they are implemented is very different and of course the data that are available for doing the research evaluation varies considerably.

Getting back to, is this program, is this initiative working, I would like to see more attention given to meta-analysis or syntheses of the research that exists.

There have been a few attempts at that. RAND has done a meta-analysis of the research on charter schools. They did a study in 2001.

Their inclusion criteria were set very high. So, only three studies were included, and they found mixed results.

Chris Nelson and myself, we did a study at the very same time, and we found 17 studies. Our bar was much lower.

What we did is, we weighted the quality of the -- we weighted the studies based on the quality of design and set up a weighting formula for that. Then Chris and I updated that analysis and published it in a chapter last year.

There are a couple of different attempts out there at synthesizing the research. To me, it also seems that, given there is so much information out there, it is important to think of systematic or more rigorous ways to look at that research.

When we look at what is happening out there, we used to have the interest groups, the advocacy groups, putting spin on research.

I think sometimes, what I am seeing a lot now, especially with the charter schools, we are even seeing the advocacy groups putting the research out and putting their own spin on their own work, and the quality is very low.

I think one of the reasons for this is because we are waiting for that definitive study, and we, the academics, we the policy makers, we are waiting for a definitive study.

So, until that time we can't give weight to these studies of lesser or weaker design. I think that with some systematic or regular attempts at synthesizing the research, we can provide good answers for policy makers.

Among these syntheses of the research, Paul Hill just came out with an analysis of the charter school research.

He basically grouped the studies in a table, positive mix and negative, and Bryan Hassell has done a

similar job at grouping the research.

What we tried to do, we tried to chart out the various studies that have occurred based on the quality of the design as well as the impact.

When we did our meta-analysis, Chris Nelson and I, we found that the overall impact rating was very slightly negative, not significant, but very slightly negative.

Of course, what we can see is that the research varies considerably from state to state. We do see like a number of the studies in Michigan, where they are relatively low on quality, and that is mostly because of the data limitations, but generally all very negative.

You can see some of our own studies that we have plotted out are in capitals and in blue, but the picture is very mixed, as one can see.

When we do these meta-analyses, it is very important, the issue of inclusion and exclusion definitions or criteria that we set for the synthesis of the research.

I know that, when I looked at the research and we aimed to do a meta-analysis, we didn't have effect sizes in these studies, as Gene Glass would have told us we had to have that.

So, we looked, well, can we do a best evidence synthesis like Flavin has advocated, and even there, the

rigor of the research wasn't in that ballpark.

This methodology here was weighting the studies based on not only the quality of the design, but the years of data available, the scope of the study -- the scope of the study defined by the number of grades covered, the number of subject areas covered, as well as the proportion of the schools that are studied.

It helped us with that approach of voiding the studies to give some quality weighting to the relative impact of those studies.

Again, the results are overall mixed at best, as we can see, but interestingly, as mixed at best I would say the body of research on charter schools appears to be, it is certainly not an evidence-based reform as yet, but over time I hope to get a better and better answer as to whether or not charter schools are working. Thanks.

PARTICIPANT: In terms of the charter school study, one term I did not hear you use was self selection. Do you think there are some differences?

In other words, one treatment would be a new math curriculum, say, for students, and another treatment would be like a charter school, but kids choose to attend or their parents choose for them to attend.

DR. MIRON: They are very strong, unfortunately, and we spell out those limitations in all of those studies

about the issue of self selection.

It is not only on the part of family self selection, but also in the schools and the particular designs that they have.

It is one of those -- you know, the level of rigor that we can apply in our studies, we can't explain all the alternative explanations, and that is one that is very hard to control for, given the data and the designs that we have used.

PARTICIPANT: You mentioned a couple of good data sets that you had with matched individual student records. I am wondering if you have addressed or would care to comment on issues related to attrition or whatever it might be in both kinds of data sets.

DR. MIRON: When we have -- if we take the Delaware data set, where we could rigorously control the students, what we are doing in the second year now is, we are trying to rule out more of the alternative explanations for the gains or losses that we see.

One of the things that we define, and we have to look more closely at is the characteristics of the stayers and leavers, and that is something we can't do with the group level data.

With this data, we can see what are the characteristics of the students who had taken test event

one, but not test event two, and we can find out what is happening to those students who are leaving.

Another possible explanation that we will examine more closely is, these designs assume the students progress a grade each year.

What we are seeing, in particular in one of the Edison schools, is a very high proportion of the students are being retained, they are not being promoted. So, they get dropped from the analysis.

This is one example where the kids, two years later they are not in the next test event. So, they are dropped from our analysis.

So, in year two of our study now we are trying to rule out more of those alternative explanations and that is another example of how we are trying to get at that.

We can do that with individual student data, but when we have group level data, we can't control for the noise of the students moving in and out.

PARTICIPANT: [Comment not caught by microphone.]

DR. MIRON: We are planning to do that with the Delaware data. We need both the matched student design as well as some of the weaker designs that we have applied in other states, like the residual gains approach.

My guess is that they are going to be very similar still, but I am very confident with the matched

student design because I can rule out more of the alternative explanations.

PARTICIPANT: On the NCLB, when schools go into program improvement, by the time they get to year three or four, there are some major changes that should take place.

There is a sort of short menu list that the federal government provides, including the elimination or the removal of the majority of the staff leading to the failure and a few other choices, including charter.

I would suspect that, with this group-level data, you won't know which action was taken, the removal of the principal or -- well, you would know the charter school, maybe, because there is a certain code or way of denoting that.

It seems that, as Michael [Scriven] was pointing out in his paper, in that situation, it almost seems to be a case by case basis to determine whether or not program A or B or action A or B was successful, when from a larger policy issue, that is what people want to know, removal of principal and staff, was that more effective or would turning it into a charter school be more effective, and are those the challenges we are going to face when we do that analysis.

DR. MIRON: It would require more data collection in the field to decode or to capture those things. Just

coming back to a point that Michael made also about the idea of cheating or inflating or deflating scores over time, one thing we did, and we couldn't control for because we didn't have individual student data for this, but in our Edison study, we looked at longitudinal trends over time.

About half of those schools were charter school studies, which is why I mentioned them today, and half of the schools in the study were contract schools, where they actually came in and took over the school.

The Edison effect, you have probably heard of what is called the Edison effect, but what we found was that the Edison effect, in year one, even when there was the same administrator, same staff, same students, because they weren't a self selection school necessarily, they were the same school, but they took them over, and there was a drastic drop in the first year test scores.

That was, in part, because of the change in management and so forth, but we also had to wonder if there wasn't something done during the testing time to lead to those low scores, because it was very advantageous having very low scores, because of some change in subsequent years.

In their contract schools, that was something we noted across the board, but we didn't have the type of data that would let us control, and we weren't allowed access to

the schools.

So, we couldn't control whether there was something else happening in the schools. It was very interesting, just one example of how schools -- also, when there is a change in a test also, it is advantageous to score low in the first year, and there are a lot of ways that administrators can work with that, and that is testing all the kids in the afternoon, you have got a larger window for testing the students, you block the students at one time.

In subsequent years, of course, you have the kids test in the mornings, in less crowded conditions and better working conditions, and test results go up.

That is one of the things that is very hard to control. Given the nature of the testing regimes and the high stakes nature, it is common practice.

Your figures seem very surprising, but it is not really surprising, in that sense, the way that schools try to alter their scores, artificially low at the beginning, or to raise them over time.

PARTICIPANT: Doesn't value added help you get at that problem, of kids moving into charter schools, or the start up of new charter schools, where some of your data is in a public school and then, as they exit, you also get public school test data after they leave the charter

school? The Hanushek study and the SUNY lab study does that, I think.

DR. MIRON: You need individual student data.

PARTICIPANT: Yes, that is what I meant by value added.

DR. MIRON: We can control for it. By controlling for previous test scores before the intervention, we can capture the type of the students, whether they are high-performing.

Demographics alone don't explain performance levels. They explain a great proportion of the variances in performance levels, but not all of it. I am sure that is possible to do.

The one state that we have been able to work with that is in Delaware. What we found, though, is because the testing event is so many -- if they were testing every year, we would have a good set of data.

What happens is -- this is when you get into the practicality of the state data analysis. What you have are schools -- the data, we link it to school variables.

What happens is, all the kids are changing schools. They are not only changing from a traditional public school to a charter public school. All kids are also changing schools when they go to grade six, when they go to middle school.

It is very difficult. Unless you have a good configuration of feeder streams on how schools feed into one another, it is very hard to control for charter school differences when everybody is moving schools in some of those grades.

DR. BARNETT: We will take a break and then start back right on time.

[Brief recess.]

DR. DUNBAR: It looks like we are succeeding in one thing, and that is getting lots of side bar conversations going.

That is all the better, because we have two sort of scheduled sessions for the remainder of the day. The first is the discussion and synthesis panel that I am about to introduce and then, immediately after that, we will have time for an open floor discussion of issues that have been discussed throughout the day.

Our plan is also to conclude the day tomorrow with a similar kind of synthesis panel for the day's presentations, and then general discussion and concluding remarks.

If you should need to leave before our presentations are over, you might want to check the weather forecast in the morning.

Stuart is going to tell us in more detail what

the contingency plans are for weather related issues and, as Bob Linn was just saying, gosh, you know, three to six inches of snow, that is just like a dusting. That is what winter is for. So, hang on for any announcements related to that.

Agenda Item: Discussion and Synthesis Panel.

DR. DUNBAR: What we are going to do now is have three individuals -- Laura Hamilton from RAND Corporation, Bob Linn from the University of Colorado, and Judy Singer from Harvard University -- each take about 30 minutes to discuss their reactions to the presentations that were given this morning, and to add any comments of their own relative to issues that have been the subject of our conversation. So, I am going to let Laura get started.

Agenda Item: Synthesis Panel.

DR. HAMILTON: It is really a pleasure to be here, and it was especially a pleasure to read these papers. I learned a lot from them and I think there is a lot of good work going on, on this topic.

I apologize in advance if my comments aren't as coherent as they should be. I found that I came here with fairly well-organized notes, but ended up scribbling in the margins and crossing things out throughout the day as topics came up or as people made points that I was going to make.

Also, I am going to try to not be redundant with Bob and Judy. So, there are certain topics that I am certain they will address. So, I will try to stay away from those.

A few times I will probably refer to a recent study that we completed that used this data base, just as another illustrative example.

So, first, I am going to talk a little bit about each of the four papers individually. It is not an exhaustive set of comments, but just a few reactions that I had for each one. Then I will try to discuss some common themes that cut across them.

First, Elizabeth Stuart's paper, I thought, if you haven't read the paper, it is a really wonderful, clear discussion of causal inference.

A lot of the material is familiar, but I don't think I have seen it presented this clearly by anyone else. I think it is the kind of description that can be made accessible to some of the less technical audiences that really need to understand this stuff.

So, there are district administrators, for example, who are making decisions about what curricula or what programs have scientific evidence to support them.

I think that she does a nice job of kind of moving beyond the idea that random assignment is the only

way, and that there are other designs that approximate that, to some degree, and explaining how they approximate that.

One of the most important contributions, I think, is that she really calls attention to the need to understand what the control condition is, and this is something I will return to a little later.

She provides a very simple example of a reading curriculum, and thinking about, when you are thinking about the effects of a reading curriculum, are you testing that against the control condition of some alternative reading curriculum, whatever reading curriculum happens to be out there, is there some alternative, no reading curriculum, and that really makes a difference for what kind of effect you would expect?

So, if you are trying to interpret a study or do a meta-analysis that combines results across studies, it is important not only to understand your treatment, but to understand differences in the control conditions across studies.

She also talks about the stable unit treatment value assumption. I don't think that was in the presentation, but it was something that she brought up in the paper, which is one of the assumptions underlying efforts to make causal inference from studies.

She notes that violations of that are common in education. She talks a lot about spill over effects that can result when members of the treatment group and the control group interact.

I think it is also important to note -- and Michael [Scriven] got at this as well -- that some of these spill over effects can occur even in the absence of interaction.

So, if you have a particular program that is implemented in certain schools, even if the control schools don't interact directly with those schools, it is possible that they might pick up aspects of that program that they think might be effective.

We saw this in our study of Edison schools, which is the recent study that used the AIR data base. There were places where schools saw the Edison program and said, well, I don't necessarily want to adopt the whole Edison model, but I like their math curriculum, or I like the way they do interim testing. So, they will pick up pieces of it, which really kind of muddies the comparison.

I think the last point I wanted to make about her paper is, she talks a lot about the need to make sure that your control and your treatment groups are as similar as possible on as many covariates as possible.

I think that is one of the limitations in the

existing data bases, that we simply don't have the kind of information that we would like to have to make those comparable groups.

Being able to link the state data with the census data can help. There is information in the SSASD data base that could be particularly useful, but that is not available for all schools.

So, thinking about some of the additional information that we would want in order to really make groups equivalent is something I will return to in the end.

Yeow Meng [Thum]'s paper, I liked the idea of describing accountability metrics as gross productivity indicators, and trying to get away from making overly strong attributions of effectiveness to schools or teachers.

I am not sure that it is feasible to communicate that adequately to users, and to prevent misuses that are an over-interpretation of it.

I think the framework of that is a nice way of thinking about what we are really getting from some of these value-added and gross measures.

My main comment with Yeow Meng's paper had to do with his call for the importance of having a valid interval or developmental scale, and I think there was already some discussion of that.

So, I probably don't need to say a whole lot more about some of the problems with developmental scales, except that there is some good research that is going on now that is looking at violations of some of the assumptions that are underlying growth models when developmental scales are used.

Joseph Martin, who is one person who is doing some of this work looking at how the sort of dimensionality and the specific skills that are measured change across grade levels.

So, when you are trying to use a developmental scale as a measure of change, you are actually not measuring the same from one point to the next, and that can actually distort.

He has actually done some simulations as well as some empirical data analysis to show how that can distort estimates of teacher school effects. I think that is one thing to keep in mind when you are using those kinds of data.

There is also a concern about the extent to which changes in a test score map directly onto changes in school or teacher effectiveness.

This is something that we looked at in the RAND study of value-added models. Dan Koretz, who was involved in that work, likes to give the example of remedial

reading. He, in a past life, actually did some teaching of remedial reading.

He gives the idea of, if he has got two sixth graders and one of them is reading at the sixth grade level and one of them is reading at the third grade level, it is really hard to achieve the amount of test grade increase for the kid who is reading at the third grade level, than it is for the sixth grade level.

So, trying to do simple comparisons and determine whether teachers and schools are equally effective when kids aren't starting out at the same point, it is problematic. It is another reason why we need to worry a lot about making sure that we match well.

Then, I think the other point I want to make about Yeow Meng's paper is, he had a nice example of using successive cohorts to try to make inferences about changes in school performance.

I think one of the problems with those types of growth models is that we are often missing data from the early grades, particularly in elementary schools where the formal testing might not start until the end of third grade, when the kids have spent sort of two thirds of the time that is going to be spent at the school, and that is where you are sort of getting your baseline.

To the extent that the school is effective in

improving achievement in the very early grades, you might get distorted growth measures if you are only looking at grades three to five.

Solutions for that might be to try to collect data earlier if possible. If that is not feasible, then I think it is important to kind of acknowledge that, if we look at growth, particularly in elementary schools, that we are looking at a limited slice of the whole elementary school period.

Michael [Scriven]'s paper, I thought, was a really nice discussion of how causal inferences are made in other scientific fields.

I get very frustrated when I hear people say, why can't we be more like medicine, and they always do randomized experiments and why can't education sort of adopt that model.

I think it is a great discussion of not only the other ways that you can make valid causal inferences, but also the problems that are associated with randomized control trials in education.

His discussion of elimination analysis, ways to sort of eliminate alternative hypotheses is, I think, a helpful way of thinking about it.

For the process to work, there are a few things that we need that we don't always have in education. There

need to be some well understood theories about relationships between treatments and outcomes.

There need to be highly detailed data. He suggested getting some of those from case studies. We can also think about improving assessment systems to get better diagnostic information about kids' strengths and weaknesses that might help us understand what is going on.

I think the importance of case studies is something that hasn't received enough attention. One of the concerns I have, based partly on personal experience, is that, when you try to get a representative sample of schools in your study to do case studies, you often find that there are certain kinds of schools that will let you in, and others that really don't want you anywhere near them, and those schools are generally not equivalent.

I think that, in addition to the expense of doing case study work, one of the challenges is trying to get reasonably representative samples of schools.

Then Gary [Miron]'s paper, I thought, provided a nice framework for classifying different study designs. I think that the suggestions for thinking about synthesizing data, or doing meta-analyses in ways that take into account the quality of the study are important.

I think some of those quality variables are pretty easy to categorize the way Gary did, but I think we

also have to think about idiosyncratic features of the studies that affect their quality that might not be so easy to characterize. So, that is one of the challenges with doing that type of synthesis.

The other thing I wanted to raise -- and he sort of beat me to the punch here -- but when you are thinking about measuring gains some of these sort of successive cohort studies that he talked about, one of the challenges that is particularly prevalent when you are looking at charter schools or other schools that start up from scratch, is that you often don't have a pretreatment baseline.

He gave the example of his Edison results, and I was going to bring that example up, and I thought it was new, and it is not.

We just published a study looking at Edison schools nationwide, and we found the same thing that Gary did on a much smaller sample that was done several years ago, which was that, if you look at Edison school performance starting from the end of the first year, which is the only baseline that you have for all Edison schools, they really improve over time, and Edison itself has made a lot out of this in their own reports on their achievement.

From the schools where we do have a pretreatment baseline -- so, for the conversion schools where we have

spring prior to Edison's entry -- what we find there is that there is a big drop during the first year.

This is something we can't measure in the start-up schools. So, we really don't know how much this would generalize to the start-up schools.

There might be something about the start-up schools that would prevent them from having this drop. We also don't have any reason to believe that it doesn't exist.

So, thinking about what the baseline is and trying to identify an adequate baseline when all you have is school-level data, I think it is sometimes difficult.

This is a case where having individual student-level data certainly would have helped us understand that first year phenomenon.

Now, let me just talk a little bit about what I think are some broad themes that cross most, or all, of these papers.

The first one is, when you are doing a comparison between treatment and control conditions, it is really important to understand both the treatment and the control conditions, what those are, and how they vary within their group.

Charter schools provide a good example. A charter school is not a sort of uniform treatment. Charter schools

vary on a number of dimensions including what kind of students they serve, the level of autonomy they have, the amount of funding they have.

They vary in the curricula that they use, their approaches to scheduling and so forth. So, trying to understand something about a charter school is difficult without also knowing something about these different dimensions on which they vary.

The Mathematica study that Gary mentioned is trying to get at some of these -- for example, they are looking at different levels of autonomy -- but there are others that are much harder to mention, but they are important to understanding what to make of the results from charter schools.

I think this issue of variation and implementation is true, even for fairly well-defined interventions like a specific reading program.

We often see, even though all teachers will nominally say they are implementing the program, we see variations in the degree to which they have actually truly bought into it, in the quality of the instruction they are providing in the context of that program, in the degree to which the school has the facilities to support it.

So, all of these things are important for understanding what that treatment condition actually is.

Then, I think the same problem applies to the control school, and it is even harder to get a handle on there, because it is often harder to get in there and get data, and there is a wider range of actions that are taking place in those control schools.

It is also important to clearly define what the control condition is. Using our Edison study as an example, one of the decisions we had to make, when we chose control schools, was whether to choose them from within the districts where Edison was operating, or to choose them from outside of the district but from within the same states.

That is an important decision. On the one hand, you might argue that, in order to keep everything equal and have equivalent groups, you need to control for district context, things like superintendent support and superintendent turnover, and all the other things that affect schools similarly within the district.

At the same time, by choosing schools within a district, we run the risk of having competitive effects that might not occur if we choose schools from outside of the district.

There are also selection effects. So, within any given district, you might assume, for example, that the superintendents are choosing the worst schools to turn over

to Edison. So, there are some selection issues there that you could deal with by choosing schools from outside the district where Edison isn't operating.

So, a very careful definition of what the control condition is, and what the threats to the validity of that are, is important.

I think this problem of finding an appropriate control group might become more serious as districts and schools become better at identifying and implementing so-called scientifically-based interventions.

So, in some sort of fantasy future, you can imagine that all districts have identified effective reading curricular and effective math curricula.

Then, when you go to test a new one, it becomes difficult to figure out, well, should we actually expect this new curriculum to perform better than the one that is currently being implemented or maybe, given that the one that is currently being implemented has been shown to be effective, all we need to show is that it performs as well.

Then finally, with this issue of understanding these different conditions, we need to take into account outside of school influences, both during the summer and during the school year, particularly when the treatment condition is designed to try to influence those outside of school conditions.

So, there are whole school reform models, for example, that try to promote more parent involvement in students' lives outside of school.

If that is the case, we need to understand the extent to which treatment effects are due to what is happening in the schools versus what is happening outside of the schools.

The second sort of broad theme that I saw when reading these papers is the importance of understanding the inferences that users are making based on results of studies, or results of accountability systems.

In the measurement world, when we talk about validity, we are always careful to say that validity is not a characteristic of the test. It is a characteristic of those inferences or decisions that are made on the basis of test scores.

I think, to some degree, that same concern applies to the kinds of accountability indices that Yeow Meng was talking about, or to some of the research studies that others were talking about.

When we present results, we need to understand what kinds of inferences users are going to make for example, about what the control condition is. What are they sort of implicitly comparing this treatment to, when they interpret that effect?

We need to do a better job than we typically do of communicating to users, what are the inferences that are actually supported by this study, and what are some inferences that might not be supported so well.

A third issue is that we need to think about that has to do with the way we test achievement is, we need to think about the degree to which the tests actually match the instruction that at least is intended to be provided under the treatment conditions.

If teachers are fully implementing the program, but the test doesn't do a good job of measuring that particular curriculum or instruction, then the results could provide misleading information.

One place this arises is when there are tracked classes. So, if you think about middle school math, for example, where some kids are taking algebra, other kids are taking a general math class, we often measure eighth grade achievement using a test that is sort of a general math test, that covers a lot of different topics.

When you do that, you may end up concluding that algebra teachers or algebra courses are less effective than the general math courses, because the tests are less sensitive to changes when the tests are more narrowly focused on one topic than others.

There are ways to deal with this, to try to break

down the data to try to understand that, but some sense of the degree to which the test is actually measuring the instruction is important.

The fourth kind of broad theme, which I think may have been covered adequately, is the problem of behavioral responses to testing, which leads to score inflation.

We have talked about this a little bit. We know that there are some ways to kind of mitigate the problem. So, don't use the same test form every year.

I was in a school once where, in California, for a while they were using the same form of the Stanford 9. When they switched to another test, one of the teachers said, "I don't know what they expect me to do. I memorized all the synonyms on the SAT-9, and now I have to memorize a new set of synonyms."

That is a sort of extreme example of what can lead to score inflation, but I don't think it is all that uncommon.

Some of these sorts of obvious steps are not necessarily enough to prevent the problem. There are some subtle things that happen.

Teachers are often aware of particular styles of items, and will coach kids on those styles. There are ways that questions are sometimes asked, and those sometimes vary by publisher.

For example, one publisher might always ask about probabilities of a spinner problem, and not ask about probability in any other ways.

Inspecting tests for those kinds of consistencies, and trying to understand how they are influencing instruction is important for kind of gauging the likelihood of score inflation.

Then we are talking about using a percent above cut or percent proficient. There is a specific type of behavioral response that is particularly of concern, and this has to do with what are often called bubble kids.

I heard the term, "bubble kids," several years ago and didn't know what it meant, and now it is almost ubiquitous.

I think you could walk into any school in the country and mention bubble kids, and the teachers will know exactly what you are talking about.

Those are the kids who are performing sort of right at the cusp of proficient. So, if your goal is to maximize the percent proficient, that is what is being reported, then a very rational response is to sort of devote attention to the kids with the greatest likelihood of achieving that, of passing that bar.

We have seen that. We have surveyed teachers and found that quite a lot of them are pulling those kids out

for separate instruction or targeting their instruction to those kids.

Sometimes there are after school programs specifically for the bubble kids. I think they call them something else, but that is essentially what it is for.

So, I think the bubble kids problem is one that is going to become increasingly salient, as we focus more on a percent above cut metric.

A related issue is the extent to which student motivation varies across these tests. We haven't talked a lot about that, but there may be different incentives that are additional factors that we need to take into account.

Then my last broad point had to do with limitations of using a percent above cut and why we should have scale scores, and I will just skip that, because I think that was adequately addressed.

So, what do we do about all this? It is easy to point out problems, and I think it is sometimes hard to identify solutions.

I think there are ways to improve the data collection, some of which are more realistic or more feasible than others.

One thing that I think we need to try to do to try to deal with some of these problems dealing with missing baselines and so forth is following individual

students over time. Judy [Singer], at least, is going to talk more about that.

I do want to mention that I have seen a lot of cases, though, where the analysis of those data is not always as high quality as it could be.

So, just having the individual student data doesn't necessarily give us valid information. Obviously, the analysis needs to follow.

It is also sometimes useful to take test scores, to the extent to which tests are developed to support this, to break them down into subscales, focused on different topics or different skills.

That can help us understand some of these issues related to test curriculum match, for example. There may be differential effects on different kinds of achievement, and we might be able to better match the instruction to the achievement if we look at the tests in that way.

Developing a data base with links to teachers can be one way to help us understand some of the variation in implementation that occurs both within and between schools.

This is difficult for political reasons. It is also difficult for practical reasons. It is sometimes easy to get a data file that links kids to teachers, and then you later find out that there is team teaching and that the kids switch classes for math. So, who you thought was the

math teacher really isn't.

To the extent that this is possible, this can facilitate better data collection about implementation, and there are lots of people out there who are working on better ways to measure implementation and instructional practices.

One district that we work with has a sort of online tool. The goal is to eventually have all the schools on a wireless network.

So, coaches or other observers could go into the class with their laptop and actually code particular teacher behaviors, and it would automatically get incorporated into a big data system that is linked to achievement data. It is pretty neat, and it is not all that far from being workable.

Then, I think that part of what would make these large-scale data bases a little more usable is good information about the testing context.

Earlier, there was some discussion of meta-data that had some of the information about the state testing system, but I would be looking for something even a little more detailed than that.

Understanding the testing context so that we can evaluate perhaps the likelihood for inflation, to what extent do these tests have the characteristics that we know

can prevent score inflation, methods for selecting cut scores or for scaling the tests, the extent to which the test is matched at least to the state standards, if not to the curriculum that is being used in each district.

So, having state profiles like that can, I think, help us make interpretations that will enhance our ability to compare results across states.

Then, of course, the last thing that I think could be done is specifically at the high school level. One of the limitations of a lot of our data bases is that we really don't have good information on high schools, in part because NCLB doesn't require high school testing to the degree that it does at other grade levels, but also because it is actually difficult to test high school achievement, given the tracking that goes on, the different courses that students are taking, and so forth, and the kind of 11th grade reading test that is often used may not tell us what we want to know about what is going on at the high school level.

Given the policy interest in improving high schools, and the need to understand what is happening at that level, better data collection for high schools is something that I would try to add, if that was at all feasible. I think that is it. I can turn it over to Bob [Linn].

Agenda Item: Synthesis Panel.

DR. LINN: Thank you, Laura. The only thing that would be worse than following one very articulate young woman would be following two of them. So, it is lucky I get to speak in between.

Anyway, I want to echo some of what Laura said in terms of complimenting the papers. I think that the four papers together provide a nice foundation, an excellent foundation, for thinking about both the theoretical and some of the methodological issues that are faced when you try to use data sets like AIR has put together or the department will be putting together, to do what Elizabeth [Stuart] was very clear about what the goal is, and that is to identify interventions that teach students more effectively.

So, this is a noble goal, and one well worth the effort that we are putting into it in talking about it, and trying to put together a data base that would facilitate what we do.

I think that Elizabeth divided these into two categories. The papers by Elizabeth Stuart and Michael Scriven had things at more to do with things at the theoretical level.

It may have something to do with age, but I was taken by the RCT zero and recalling, back in the days of

the follow-through experiments and efforts to have randomized control treatments that are then going to be implemented in real classrooms with real teachers and real kids and real principals, and they all have as their goal to undermine the experiment.

Their goal is not really to go along with the experiments, but to teach kids more effectively, and that is the goal of the researcher, but it is also the goal of the teachers, I think.

So, if they see something that is working in another place, why wouldn't they say, well, I can try that, too.

In the Edison example, if you see things that Edison is doing that you think look good and effective, why not pursue them on your own.

On the other hand, I was also taken with Elizabeth laying out what I think of as -- the first paper where I really thought in these terms was the one by Don Rubin that she relies upon a fair amount, to lay out this, what is it that we want to know.

We want to be able to observe everyone under both conditions. We can't do that. We can only observe them under one of the two conditions, and random assignment, if you can make it an RCT-2, it is obviously a desirable way to do that, because it makes you less dependent on being

able to find all the right covariates to do the matching or to do the adjustments.

I don't think there is any quarrel there between what Michael said and what Elizabeth said, really. If you are talking about RCT that are the two, the double blind, as opposed or the RCT-1 or the RCT-0, which is harder for us to do in the field of education, a double blind, than it is in some other fields. I am not saying it is easy in other fields, but at least it is more doable than it is in ours.

Yeow Meng [Thum]'s paper, I thought, moved into more detailed discussion of a particular approach to analysis, at both modeling or value-added, or what he was calling today a value-added hypothesis, I guess, of trying to tease out effects that you associate with schools, in the most part, for value-added modeling the way that it has been used. I will say more about that in a minute.

It is not so clear exactly how this applies to the data base issue because, for the value-added modeling, we do need the individual student data.

Incidentally, for many of the things that Gary [Miron] wanted to do, with the charter school analyses, he also needed the individual student data.

So, it is not arguing that you can't do anything with the school-level data, that there are a lot of

purposes for which there is really no substitute for getting at the individual student data.

Gary strikes me as a brave researcher, to go into this field. I can't think of many areas of research that are going to be more controversial than looking at charter schools.

One of the things that he has going, to his credit, is that as he pointed out, and as you could see in his chart, the blue dots were, in some cases, quite favorable, so that the opponents were throwing darts at the study.

I am sure that, in other cases, they didn't look so positive. So, it was the proponents that were throwing the darts.

Years ago, I was on the test standards committee when Melvin Novick was chairing it. He used to lecture, as only Mel could, the joint committee that was revising the standards and say, we will know we have got it right when the heat from the left is equal to the heat from the right. That is kind of maybe where Gary has managed to come out on this.

So, I think it is a tribute to the objectivity of the studies that he has managed to carry out in the area of charter schools.

Elizabeth is very clear on saying that she thinks

the school-level data base can be used to tease out causal interpretations of the type that she was saying is the goal.

She wisely cautions that this has to be done with considerable care. That also, I think, would apply as the RCT-1, 2 and 0 comments apply, that you can't do randomized control trials without a lot of care. So, this applies well to this particular data base, as well as other places.

The goal, as she lays it out, is to try to replicate a randomized experiment when you are dealing with something like this data base.

What that implies is that you need to be able to get a lot of information about schools, so that you can do the kind of matching that she proposes is needed in order to satisfy Don Rubin's notion of what is a strongly ignorable treatment assignment.

Strongly ignorable treatment assignment is what you need if you are going to take the kind of comparison group, and be able to say that the differences you observe are attributable to something other than the treatment assignment.

This requires that you have to have the assignment be done in a way that is independent of the outcomes, given the observed covariates, and that you have a positive probability of receiving each treatment, the

treatment or the control, for all of the levels of the covariates.

Now, that is a challenge for many of the things that we want to study. By design, treatments are often targeted toward a certain end of the distribution. They may be targeted only toward children who come from low income families, which makes it harder to find a matching group that will not be receiving that treatment condition.

That suggests the need for not only a great deal of care, but probably more information than is available. This has come up several times today, the notion that the data base, as good as it is, is not going to have all the information that people are going to want, when they come at looking at these causal interpretations of particular treatments versus alternative conditions.

Another issue that Elizabeth discusses is the stability or the stable unit treatment. This we also heard quite a bit about, the problems that you have with possible spill over effects, the leakage, as Michael [Scriven] was telling in his paper.

In addition, you have the concern that you want to have a treatment that is really one version, and you want to have a control that is one version.

That is a rare thing in education, because you can give everyone so many milligrams of the drug, but it is

not as easy to give them so many minutes of a particular type of instruction.

So, what you have, indeed, is a great deal -- many versions that get labeled the same thing. So, you have a whole bunch of variations on the treatments that were put in place going back to the follow through experiment.

One of the important things that came out of the follow-through experience, other than learning how hard experiments were to do, was, the realization that you really needed to have implementation studies.

This harkens to some of the comments that were made about finding out more information on the ground about what is really going on.

Once the implementation studies get underway, being run by SRI at the time, the Stanford Research Institute, it soon became apparent that, just because something was called the Engleman method did not mean that it wasn't also a bunch of other methods, or the non-Engleman method wasn't also partially the Engleman method, because the implementation just varied so much all over the place.

Charter schools are another nice example of the fact that you have all kinds of variation. Here you have variation somewhat by design or by law. There are certain types of charter schools.

I was also wondering, in looking at figure one of Gary [Miron]'s [paper], the degree to which the different locations were, in fact, different meanings of charter school as it related to state.

So, it may be that it isn't just that some of the studies are finding that the charter schools are more effective or less effective than others, but that which state they happen to be in, the charter school itself may be quite a different thing, and how kids get into the charter school.

So, there is a potential confounding there that I think would be interesting in the meta-analyses to try to tease out, if you could get more information about the laws and how charter schools came about in the various states.

The second challenge is -- well, not the second challenge, but there is the gigantic challenge of trying to find a comparison group that is good enough, similar enough.

First, there is the problem of finding them. If you had treatments that were targeted toward only one end of the distribution, that often leads to techniques that, on the surface, seem like they might work, like a regression technique, but the example of looking at regression functions that are limited to a certain part of the distribution and extrapolating, you don't have any

great confidence once you start extrapolating to another part of the distribution.

A second problem is obtaining enough information. That harkens back to the need and a suggestion that the data base needs to think more expansively in what kind of information can be included about schools, so that, with better information about schools, better matching would at least be a possibility.

I want to go back, then, to Yeow Meng [Thum]'s discussion of the value-added model. The value-added modeling or hypothesis generation, however way you want to call it, is certainly a topic that, while it may not apply to the data base, is one that there is a huge amount of interest around the country.

A few states have laws that you have to use the value-added modeling as part of the accountability system. Tennessee has had that for quite some time. Ohio has a law that has been put in place, that part of their accountability system has to look at it.

On November 21 of this year, Secretary Spellings announced the pilot program that invites states to submit proposals for growth models using an approach as a way of assessing AYP for the No Child Left Behind evaluations.

It is worth noting, however, in her letter that she lists -- the letter went to the two state school

officers, and it listed seven core principles that these growth models have to abide by.

The first one, when I first read it, I thought, well, this kind of undermines the whole idea of the growth modeling, and that is the not giving up on the 100 percent proficient by 2014.

So, the growth model has to get you from where you are to 100 percent one way or the other. So, that is going to put a real crimp on the system, I think, of having a value-model approach to AYP that is sensible.

Now, I shouldn't say that, maybe, in this building, but I have been saying it all over the country, and I might as well, and that is this idea that we are going to have 100 percent of these children proficient by 2014 is just completely ridiculous.

So, if you have a condition that you have to have something that is completely ridiculous, then how can you have a reasonable system of using growth models as an approximation for AYP. That has nothing to do with the data base. So, I won't rant on that any more.

DR. DUNBAR: There will be time for questions.

DR. LINN: The value-added, the label, value-added, is a terminology that invites a causal interpretation, very clearly.

You don't say value-added by the school or by the

teacher without meaning that it is because of something the teacher did or that the school did, that adds the value.

Nonetheless, I am very sympathetic to Yeow Meng's caution about using that causal interpretation, and he is not alone in that regard, as he referred to papers by Steve Raudenbush and Don Rubin, that are also warning that the value-added modeling, because it can't rule out many of these other alternative explanations, one of which Laura talked about, the differences in first grade when most of the systems are going to start with a baseline year of third grade, or possibly second grade.

Kids come to school with huge differences at the beginning of kindergarten, and beginning of first grade, and those may have important implications that need to be ruled out as plausible explanations.

There are also the possible explanations of differences in parental support, differences in socioeconomic background. If you have a system that does not include socioeconomic background in the value-added model, now you can include it, as it is done, say, in Dallas, but not in Tennessee.

Analyses by Sanders and Ballou have suggested that controlling for the background characteristics doesn't really change things much from the usual background, like free and reduced lunch, doesn't change things much from

what you have when you have the value added system with just the prior test scores.

However, you can't logically rule out that as another explanation, because there are things going on in homes of kids that are not included in the kinds of controls that you are looking at when you look at that value added.

Now, having said that, I am in favor of the value-added modeling. Let me say that because I think a descriptive measure will get you a lot closer to having something that, in Yeow Meng's terminology, at least has hypotheses of things that you could then go in at a second level, if you are thinking of an accountability system of really triggering a closer look, as opposed to automatic sanctions.

A system like that could use value-added modeling very well, I think, to target where it is that you ought to go to find out more information.

I think I already commented on things that I was going to say later about Gary's paper. The other comments I have are redundant because they are really talking about the percent above cut, except in one regard.

The percent above cut is going to be important in the data bases, to have a better understanding of how strange the cuts are in different states.

The stringency of the cut scores in different states varies all over the map. So, there are a very small number of states that are actually more stringent than NAEP, at least looking at it in the norm-referenced sense of how many kids get above the proficient level.

Most of them are somewhat more lenient and some of them are a whole lot more lenient. There is also the issue, then, of states that have redefined their standards.

States that may have set standards pre-NCLB tended to be very stringent. States that set them post-NCLB tend to be more lenient. States that have reset them since NCLB also tend to be more lenient.

So, at least having an understanding in the crude sense of how stringent the state standards are is going to be important for using the data base, and keeping straight that -- for example, in Colorado, when we say proficient for state purposes, we mean one thing. When we say proficient for NCLB, we really mean partially proficient.

So, there are those issues that need to be addressed. If you are looking at closing the gap, a paper a few years by Paul Holland and some others, and some more recent work by Don McLaughlin, shows how big the gap is.

There was a comment earlier this morning, too, about the gap is obviously going to depend on where the standard is set.

You could have states that looked like they were closing the gap if you were looking at the basic level, that looked like the gap was expanding if you were looking at the proficient level or vice versa.

So, understanding more about where those cut scores actually are, I think, is important. I will stop there, and turn it over to this young lady.

Agenda Item: Synthesis Panel.

DR. SINGER: Thank you all for staying. That is the first thing. Batting third discussant after a day like today, I find it very hard to think of some comments that my colleagues wouldn't say. So, I am going to try not to be redundant.

I am going to start off with a reflection that is from someone I now know is from the mid-20th Century. I have learned more about age today than I thought I would.

In 1962, John Tukey published a fantastic paper in the Annals of Mathematical Statistics entitled "The Future of Data Analysis."

Now, publishing this in the Annals of Mathematical Statistics was, in and of itself, an interesting exercise. Here is a guy who is really a mathematical statistician.

He wrote this paper. In it, he begins to presage his later work in exploratory data analysis. It is fun to

read in retrospect.

In particular, I think there are several things of relevance for today's discussion in this paper. Tukey mused broadly about what he called the tools and attitudes that professional statisticians and data analysts who tackle real world problems need to do their work.

There are two quotes from that paper that I want to read to you, one by John Tukey himself, which has actually become pretty well known, and another that he attributes to his colleague, Martin Wilke, which I think has languished in obscurity, but they are both appropriate for today.

The first is Tukey's quote: "Far better an approximate answer to the right question than an exact answer to the wrong question."

The second is Martin Wilke's quote: "The hallmark of good science is that it uses model in theory but never believes either of them."

I first read Tukey's paper in graduate school, and I invoked that wisdom about better a sort of approximate answer to a good question, whenever people come to me asking, how can they gain insight into questions of deep importance to them, using data that are flawed, at best, and were often collected for other purposes.

As I was reading the materials to prepare for

today, I found myself thinking back to that paper, because I think that is exactly the situation that we are in here.

It gave me great solace, actually, to go back and read it, because the original first sentence in my remarks today is, is there any way that I can be optimistic. Is the data base so profoundly flawed that nothing is possible?

So, I sort of said that is not a very productive way to spend my time, and I didn't think it was a productive way to spend your time.

Instead, I am recasting my remarks and offering them in testament to John Tukey, and Fred Mosteller, who I hoped would actually be here today by saying, What would John Tukey have said in reaction to today's paper?

I have little doubt that he would want us to be very constructive and forward thinking, instead of spending a lot of time thinking about what is wrong here.

So, with the strong realities of what the data set actually has, I would like to offer 10 ideas that might be food for thought for improving the current data set, or improving the kind of statistical models that people use with the data.

Because I knew that both Bob [Linn] and Laura [Hamilton] would say a lot about measurement issues, I am going to say virtually nothing about them.

I am not making light of them at all. In fact, I

would say that measurement issues probably trump just about everything that I am about to say, but recognizing that the data set has those issues, I would like to talk about two broad classes of ideas.

Half of them are going to focus on improving the statistical models that could be used, and the other half are going to focus on additional analyses or additional data sources that might be added to the data base itself to make it useful.

The lines between these categories are not sharp at all. So, some of these things may be categorized incorrectly, but I am hoping that, taken together, it is not that these provide answers, but I think that they stimulate conversations, and they get people thinking about some way that these data could be used productively.

So, let me start out with five issues of statistical modeling. Number one, let's be sure that our statistical models are carefully and fully specified.

Elizabeth Stuart wisely makes the case that researchers wanting to make causal inferences from this data base need to be much more explicit about their underlying causal model.

I would extend this point much more broadly than just the underlying causal model, and say that I think researchers who are using these models need to be much more

explicit about which models they are using.

Here, I am not trying to pick on anybody in particular, but I did take the time to read the background papers, and they vary in the degree to which models are clearly specified.

When I talk about specifying the models, it is not enough to state an analytical approach. It is not enough to say, "I will use regression," or it is not enough to say, "I will use multilevel modeling," or some buzz word. That is not the point. That is not the model clearly specified.

The same thing about a comparison group. It is not enough to say, "this is how I am going to construct a comparison group."

What is the underlying model here? After all, when you use statistical methods to analyze data, you are fitting models to data, and those models are, in fact, your representation of what the model looks like.

You are going to use sample data to estimate the parameters of models, but you need to be very clear about what those models are.

When you specify the models, it is not enough to specify the structural portion of the models. I think the papers that we read in preparation for this gave a lot of attention to the structural portions of the models, the

effects of covariates, for example.

There also needs to be a lot more attention to what is called the stochastic portion of the model and, in particular, the error structures.

This is not easy stuff. I am not saying this is something where there are off-the-shelf solutions, and lots of buzz words are thrown around, for example, "value-added analyses" or "value-added hypotheses."

I would say, not only are there not value-added models or value-added hypotheses, there are statistical models that have parameters.

You put assumptions in about the behavior of certain features of those models and you ask, "Can I use available data to estimate those parameters to draw inferences?" This is much more elaborated than I think the things going on today.

Sort of at this point in the history of educational research, I think that to say, if you didn't have data that you were starting with -- and that is really what I am arguing, don't start from the data, start from the model -- the appropriate statistical model for educational phenomena starts with the student.

I mean, wouldn't most of us think about what happens to the student when we are looking at student achievement?

It recognizes the role of the teacher. It recognizes the role of the students' classmates. Only then does it build to the school level or the school district level.

Notice how I am proceeding totally backward to what most of the models we are talking about today have specified.

There was some mention in the papers and also some mention today of what I thought was actually a dead issue, and this is the unit of analysis problem.

I actually wrote my qualifying paper in graduate school on the unit of analysis problem. I thought it had been put to rest, but whenever I come to a meeting like this, it sort of comes out, sticks out its ugly little head.

It is not an issue of running the model this way and running the model that way. In other words, it is not an issue of fitting the model to the individual student-level data, fitting the model to the classroom- or school-level data, and checking to see whether we got to see the same answer.

That is a dead question. The models are addressing different questions, and if we reframe our questions that way, we are ignoring decades of progress in statistical modeling and how we think about educational

data.

I think the limitations of the data base cause people to do it, but I think real progress could be made if we take a step back and ask, what kind of models do we want to use, and then where could we get the data to support the feeding of those models.

To some extent, we have got a missing data problem here, not unlike what is now called the fundamental problem of causal inference.

As Elizabeth said, the fundamental problem of causal inference is a missing data problem. You either observe the outcome in the treated group or the outcome in the comparison or control group, depending on how you design those things, but you don't observe both things at the same time.

Well, we have the same problem here. We have problems with missing data. We have data at the aggregate level, but what we really want is the student-level data.

I would say that what we need to start doing at the outset is specifying the models we would like to fit, and then asking whether we have the data to estimate the parameters of those models.

Starting from this, I am hoping we will begin to clarify what kinds of inferences we can draw from data and what inferences we can't.

Whether they are causal or not I am not even talking about here. I am merely talking about what kind of models would fit.

Now, as a corollary to this, I would say that we need to start fully understanding the consequences of omitting one or more levels from our analysis.

In essence, the problem we want to specify here is the multilevel model, or mix effects model as Yeow Meng [Thum] is calling it, and we have got omitted levels.

Now, there has been a lot of work recently on the consequence of omitting levels of analysis from multilevel data problems.

Most of that work focuses on the omission of upper level problems. So, in other words, you have data on students. Let's say you have data on students over time.

So, time at level one, students at level two, classes or teachers at level three, schools at level four, school district at level five, and your computer program crashes because you have too many levels and too sparse data.

So, there has been work done on what happens if you set aside the school district from that model, what happens if you set aside the school from that model, what happens if you set aside the class from that model.

I would say we have a similar sort of problem. We

are setting aside time, student and class when we are fitting these models to data. That has consequences.

I think that in order to make progress in what we can do with data sets like this, we have to think very hard about what are the consequences of ignoring lower levels from what are multilevel models, whereas most people would thinking what are the consequences of omitting higher levels from multilevel models.

The third point I want to make is actually something more philosophical. What is a school? That may seem sort of pejorative, a stupid question to be asking.

If we are conducting analyses using a school-level data set, then our -- now I am going to invoke the language in vogue -- our unit of analysis is the school.

Well, if that is the unit of analysis, what is it? I mean, is it the bricks and mortar? Is it the principal who is at the head of it? There is only one of those in most places. Is it the agglomeration of the teachers, the agglomeration of the students?

If we don't know what a school is, in other words, if that unit doesn't have some conceptual meaning to us as something that you can really get your head around about how you are studying how it might change over time, then why are we doing these studies in the first place?

There are reasonable questions about, are the

characteristics of schools that we can consider to be reasonably stable that we might want to track them over time.

Principal characteristics is an easy one. School size would be another one. Student-faculty ratio might be a third one.

If there are substantively interesting outcomes that are appropriately measured at the school level, then building models at that level makes sense.

If there are not appropriate outcomes that can be measured at that level, then I am not sure that building models at that level, ignoring all the things that go into it, will actually live up to what we are trying to do here.

Elizabeth alluded to the ecologic fact fallacy, and I think it is worth repeating here. It was named in 1950 by Will Robinson, and the example that Elizabeth gave was exactly the example that he used in 1950, about census tract data, and the percent foreign born and whether or not that ecological correlation, which is positive, makes sense to hold at the individual level.

That is from 50 years ago. It hasn't gone away. In fact, every single analysis of these data bases that is done solely at the school level, you are not analyzing student-level outcomes.

We can call it student-level data, but it is -- I

am being a little bit extreme here. I am being provocative. But you don't have the data that you think you do.

Fourth, let's critically evaluate our assumptions and conduct thorough sensitivity analysis. What assumptions are we willing to make, and what assumptions are we willing to say, oh, that is not a big deal if it is violated.

If we do sensitivity analysis, it begins to answer the question that, if you change the assumptions, the findings would still hold.

Notice I am not saying changing the models and I am not saying changing the levels. I am saying changing the assumptions underlying the models.

So, one example that both Laura [Hamilton] and Bob [Linn] talked about is this stable unit treatment value assumption, the sort of spill over effects.

It makes little sense for education researchers to invoke it. It doesn't hold. Let's just say it doesn't hold. Let's just be honest about it and move on.

Then, given that it doesn't hold, how can we try to place bounds on what are the consequences of violating it.

What is interesting in the discussion today is, we have talked a lot about student achievement data. We have talked a lot about the notion of moving from cut

scores to scale scores.

We have yet to talk about some other features of these scores, like the variance within schools, or another word that we haven't talked about at all is the interclass correlation of these scores.

Fundamentally, the interclass correlation, which measures the degree of similarity of students within a school, is the parameter that is going to tell you about what the consequences are of ignoring the lower levels of analysis in your work.

If that interclass correlation is one, go with the school data. You have everybody's blessing, because the data for one student is the same as the data for all students.

If the interclass correlation is zero, then I would ask what is going on in the school. There is no similarity there, either before -- well, it is an unusual neighborhood -- and after, well, what is going on in the school if there hasn't been a sort of bringing together of people.

So, I think we need to think that, if we can't get individual-level data -- I will talk a little bit more about that at the end of my remarks -- then we at least should start trying to get information on variability in schools and the interclass correlations that are going on

here.

Only with that information can we really evaluate the validity of the kinds of assumptions we are making when we do the analysis at the school level.

My fifth point here, something that hasn't come up, but it is in several of the background papers, is the perils of standardization.

Several of the background papers were faced with a very intractable problem, which is non-equivalent outcomes, non-equivalent predictors, and a deep desire to do the analysis across broader classes.

They end up standardizing either the outcomes, striking the mean -- either standardizing the outcomes, the predictors or both.

Now, I understand the desire for standardization. It seems so right. You have got these measures, they are non-equivalent. Let's make them -- quote -- equivalent by standardizing them.

Obviously, standardization makes them equivalent. That is what the word says it does. Unfortunately, it is not that easy.

Standardization of these measures doesn't render them equivalent, and standardization doesn't help you do some of the other things that people think it helps you do.

It doesn't help you identify the relative

importance of predictors. In fact, it is not that easy to identify the relative importance of predictors, and standardization doesn't do much to help. I have more references on this, if you are interested in it. It happens to be one of my pet peeves. So, I have spent some time on it.

The other line of reasoning is that standardization allows you to compare samples that are different from each other.

Well, in fact, standardization can do just the opposite, because if the standard deviations differ across the samples, and you standardized, and you even have measures that are equivalent to start out with, by standardizing, you have taken these very equivalent, let's say, regression coefficients, and rendered them non-equivalent, because you have divided by standard deviations that were not equivalent.

When you are in the longitudinal context -- and that is also where this comes up, because you have got measures that are not equatable over time -- you are doing it even more damage.

You are, in essence, taking away the very nature of how kids change over time. When you look at Yeow Meng's class and a lot of those had those fan spreads, when you look at test data, fan spreads are common.

If you standardize, you get rid of the fan spreads. You think you are doing the right thing, you are standardizing, it sounds good, but in fact, you are changing the properties of the measurements that you are working with.

I think it is very important to say that, if the data are flawed, there are things that you can try, but that is absolutely not one of them. It is not going to get you where you want to go.

Now I am going to turn to some additional types of analysis for data that we might collect, to enhance the value of the data base, and I am going to return to John Tukey and my first comments here. See, if you hide behind somebody else, that will actually make it kosher.

Let's appreciate the role of descriptive analysis. There is great value in describing things. In his 1962 paper, Tukey noted: We need to face up to the need for both indication and conclusion in the same analysis.

He later went on to say, we need both exploratory and confirmatory -- that is how he changed the words, but initially he said, both indication and conclusion in the same analysis.

It may be heresy to say these days, and it may be heresy to say in this building, not every statistical analysis needs to lead to causal inference.

You know, don't get me wrong, I support all of the push toward causal inference. It is a much needed corrective, I think. There were decades when educators really dismissed the kinds of work that needed to be done conducting randomized trials and studies that permit causal attribution.

Just as epidemiologists have learned a lot, especially from the analysis of longitudinal data on individuals -- and I will come back to that -- researchers can use these data to generate interesting descriptive information that would be invaluable.

They are invaluable for hypothesis generation. That is a worthy activity that should be supported. They are invaluable for the design of future intervention studies.

We should be celebrating the efforts that have gone into creating a fully representative data base. There are a little bit of missing data, but we will set that aside.

This is an accounting of every school in this country. Surely a descriptive analysis or even what I would call a relational analysis -- nothing claiming causal attribution -- could have value.

So, let's not get so caught up in everything has to be causal to say there isn't any value in plain old

description.

Now, as I say that, I think there are probably some interesting things that we could be using in a data set like this.

One of them that has not come up in any of the discussions here, or the papers that I read, or at least that I could find, is the possibility that we could use the data base to identify and evaluate the effects of natural experiments.

For those of you who are not familiar with this literature, economists, primarily, have made great strides in recent years by using extant data -- in other words, data collected for administrative purposes, like this data set -- do you know how to pronounce this? That is the other thing. We have changed the SSASD -- it just doesn't trip off my tongue here.

The natural experiments are really a compelling opportunity here, and I suspect there are lots of interesting ones buried in these data.

If people who have access to the data, and know about the phenomena, policies, that are going on in the states that they are working with, there are some great opportunities here.

In terms of the kinds of natural experiments that people think about, there are sort of the forces of nature

natural experiments.

So, there is some exogenous force, like Hurricane Katrina is the most recent example, or differences in government policies, which is probably the most profitable in terms of this data base.

For example, there are different eligibility levels across states or within states that would create the opportunity in which this exogenous variation, variation that is not coming from the individuals or the schools themselves, but from some other source, and it is providing a treatment mechanism, treatment assignment mechanism, that is not entirely self-selected.

This approach has been used with great profit to study the effects of things that you really can't randomize, like the number of years of schooling, or class size, which you could randomize, but using discontinuities that arise in class sizes as there are certain rules that say that in school size growth, there are certain rules that say you can't have classes of more than 25 or 20 or whatever.

If your enrollment grows, all of a sudden classes get a lot smaller than that, because you have got to create more classes to meet the requirements.

I would urge the people sitting in this room who are much more knowledgeable about this data base, and about

the policies going on in states that they are working with, buried in these are some fascinating studies.

The methods are not without problems. Using instrumental variables -- I could go into that, but just sticking in your head -- if I do nothing else today, stick in your head the idea that in places where policies are changing over time in interesting ways that you know about, here is a data set that could actually begin to help you tease out the effects of those policy changes.

Eight, can we use the data set to document the changing composition of our schools. Most of the analyses of these data look at changes in outcome data.

Well, there is a wealth of predictive data in "them thar hills," as they say. In fact, with future editions, from what I understand from the conversation today, there may, in fact, be more predictive data there.

Don't underestimate the value of looking at how the composition of our schools is changing over time. There is information about changing demographics, information about changing teaching forces, changing class size and other attributes of the schools and school districts.

Here I am not trying to make causal inferences. I am trying to actually understand how the schools themselves are changing over time.

I think this is a real resource that is being

underutilized because everyone is focused on outcomes. In fact, we need to understand better how the schools themselves are changing.

In other words, how do these covariates change over time, so we could then model the relationship between those time variant covariates and those outcomes that we care so deeply about?

Nine is an idea that I am not sure whether it was answered today. So, I will throw it out here, and if you tell me that it was, I will go past it.

It is this idea of using the data base to provide a comprehensive inventory of interventions being conducted in schools.

As kindly constructed, or at least as the analyses that I have read are being framed, people talk about a study of a particular -- and I will use the term treatment rather loosely -- they look at the effects of charter schools, in Gary's example.

In fact, there are lots of different treatments being applied to schools around this country. It would seem to me that this data base could serve as a very valuable inventory of what is being done to schools.

It would allow us to move beyond the sort of "I want to compare charter schools to non-charter schools," or "I want to compare reading first schools to non-reading

first schools," to understand the panoply of what is going on, both funded at the federal level -- because there is a lot of stuff the Department of Education is doing -- but let's remember there are lots of things being done to schools and school districts that the Department of Education is not involved in.

We have got private philanthropies doing experiments, we have got university researchers, we have got not for profits doing all sorts of things for schools -- breaking them up into small schools, doing all sorts of changes.

We have got for profits, we have got textbook publishers. There is a lot going on in the schools, and especially in the very small number of very large districts there is an awful lot going on.

It would seem to me that this data base provides an opportunity to start inventorying this, so that it is not in the province of a few people who get access to the identities of the schools with the different properties but, rather more broadly, let's characterize what the schools are experiencing, and let's make that information publicly available.

Now, in the spirit of all top ten lists, let's save the most important for last, it will come as no surprise -- and Laura alluded to this -- that I am going to

take my last few minutes to make a desperate plea.

As someone who has no financial investment in this data set, other than the fact that I pay my taxes, and I don't even use this data set, to put political pressure wherever it needs to go to start collecting data on individual students and teachers.

I think that the purpose of this meeting is to have people who work with the data and really toil in the field talk to people like myself, who don't work with the data, but are interested in education broadly, and think about ways that we can really make a difference.

I think this is a place where we should be talking very explicitly about partnering, about saying, if we are going to understand what is going on in education, if we want to make these causal inferences, if we want to track changes over time, if we want to do the kind of work that the department cares about and we, as education researchers care about, we need to start putting pressure on creating longitudinal data sets on individuals tracked over time.

Yes, that means creating ID numbers that could track kids. I took a cab over with Geno [Flores] over here, and we were talking about the state of California.

You know, is it possible to create ID numbers, so that you could track kids within a district across

districts, and then, heaven forbid, into other states?

If you look at other fields -- I think it is both profitable and dangerous to look at other fields. Everyone is sick and tired of the medical research analogy, give a pill, we all come up with the reasons why that doesn't work. I could argue with you and say surgical interventions aren't blind either, but they are still randomized. So, let's not all say, it is just a matter of giving a pill.

This is a case where other fields have made great strides because of the availability of longitudinal individual data, whether those data come from social security earnings records, that is what allows people to do work in labor economics, to look at the effects of all sorts of interventions, and a lot of education related predictors, like getting a GED, for example.

That is possible, because we can use social security records to link earnings over time, and then we can go back in student loan data bases and try to look at those phenomena.

In public health and medicine, the strides that have been made by being able to look at Medicaid claims data are unbelievable, because they not only talk about what happens in terms of expenditures, which is why the data bases are created, but lots of valuable work about the effects of health interventions, about the effects of

cardiac care, for example, as a result of looking at what is possible through Medicaid data bases.

I think that it is frustrating for everybody in this room, because I think this is something we all share - - I am speaking to the converted here, I don't think there is a person here who wouldn't want to have the individual-level data -- but what I am saying is, this is an opportunity to come together in a very sort of research-oriented, policy-oriented -- that we could actually answer questions that policy makers want the answer to, if we got the individual-level data.

I would put it out as a challenge to everybody in this room, and to our discussion about how we can put aside differences of opinion about, is it this model or is it that model, and is it this comparison group or is it that comparison group, because those are niceities.

The Big Kahuna is, can we get individual-level data. Are there some state representatives in this room -- and I will look at my former student, Mitch Chester -- are there some state representatives in this room who would be willing, on a voluntary basis, to say, I will work with you in trying to put together the systems that would allow you to answer the questions?

If we do that, we can begin to show evidence that is useful for me and my state -- which is why I am willing

to front it, useful for me and my state. Then, if we do it as an existence proof, it might be possible to move forward.

So, I am going to stop here and say that I have great hope that the fact that you are all here at 4:00 o'clock on a pre-snowy Washington D.C. afternoon, that you care about these issues, and I think it really is possible to make a difference.

I hope we are able to now make some progress in thinking about what is possible about these data, because we do have very important questions that we want to answer, and it is important that we take the time to try to answer them well. Thank you.

[Applause.]

DR. DUNBAR: I can't think of a better way to end a discussion panel that would encourage questions and further discussion of some of these issues.

So, I am just going to open up the floor for discussion, first of all, for the panel on any of their comments, and then more generally, the issues that we have been discussing all day.

Agenda Item: General Discussion.

PARTICIPANT: A couple of the panelists have raised the issue of test scores and changes over time and test score inflation. I wondered if people have given any

thought to possibly using this data base to assess that whole issue.

Just to provide a little bit of context, the people publishing, I guess, a couple of articles in the New York Times and a couple of other places where the recent NAEP results, the proportions proficient in the state have been compared with proportions found in NAEP -- how do I say this -- the proportions proficient found in the NAEP scale differ from the proportion proficient found on the state test, which Bob Linn has been pointing out for some time.

There has been some concern that we are seeing a race to the bottom, that states are diluting their standards on and so forth.

I guess the question is, in general, have people thought about using this data base to investigate that. I don't know whether it is appropriate to try to include Don McLaughlin now in that conversation, too, since he sort of looked at the link there at one point in time.

I don't know whether, looking at that link again at a second point in time, would help to address that or not.

DR. MCLAUGHLIN: That comparison depends on the equating of psychometrics. If you assume that NAEP's psychometric linking from one assessment to the next is

perfect, then it is fairly straightforward to do the linkage one year and the linkage the next year, or the next NAEP, and see the extent to which the state's test has migrated, or their scores have migrated upward compared to what NAEP says they are achievement change has been.

That is based on the assumptions that the equating of one year to the next on some test you can believe is completely accurate.

DR. LINN: I was going to involve Don if you didn't anyway. I do think there are a lot of questions about the linkage of state assessments to NAEP.

There are concerns that the conditions of administration are quite different in state assessments than they are in NAEP and, therefore, the motivation may be different.

I think the kinds of comparisons, especially as we get more time now, so that you can do it not just at one point in time, but at several points of time, will be very informative at a descriptive level to some of the changes that are going on.

There has been work in the past that has tried to compare an equating or a linkage of new state tests to NAEP. Even when you use the same test in different states, you find that you get different functions.

David [Thissen] is going to talk about this

tomorrow, actually. If you look over time, the stability over time is not great.

The fact that the stability over time is not great is informative in and of itself, however, because that may be giving you some hints, at least, about places where score inflation is of particular concern.

PARTICIPANT: Not to beat a dead horse, but I think for understanding score inflation, again, it is important to go beyond the percent proficient, and to have information about the whole distribution.

Your sense of how much inflation has occurred will depend a lot on how you look at the numbers, and whether you are actually looking at movement in the whole distribution, or whether you are looking at the movement above a certain cut point. I think it is sort of one more reason why we would like to have that information.

PARTICIPANT: A couple of observations, one on this panel. Pointing out the issue of fidelity of implementation in looking at these studies -- I am with Houghton Mifflin -- and how textbooks, for example, are used, how the interventions are used is a very significant variable that doesn't always get identified.

The other point that is raised was to focus on the inferences that are being drawn regarding causation or whatever.

The problem that I see is that somebody draws an inference that there is causation, that it is causative. I guess the challenge is to, in effect, provide some standards for inferences that should be drawn.

It is an issue that has been endemic with NAEP and with state testing programs. What kind of guidance might the federal government provide, and is it workable, and how good is good enough in drawing inferences about causation or whatever else. It is more of an observation than a direct question.

PARTICIPANT: I didn't want to insult you by saying, what is the question, but since you said it, that is good.

PARTICIPANT: There has been a fair amount of discussion about some of the limitations of using the school-level data as outcome measures.

There has been much less discussion, sort of, about the population of federal programs whose effects you might want to evaluate.

One of the things that, if we are going to have data on implementation, and it needs to be available and somehow linkable to the data base, if not within it already, what kinds of things do we want to have in there to answer, or to be ready to answer, some kinds of questions?

So, Elizabeth threw out three or four examples at the beginning that were fairly provocative, of important policy questions.

Just, for example, the extent to which increased funding for school libraries increases literacy levels. I would point out that that, as with many federal programs, is not a student-level intervention. It is a school-level or higher intervention.

So, using school level or higher data to try to assess potential impact seems a reasonable place to be looking.

I wonder if we could, as we are thinking, maybe in the morning tomorrow or thinking about ways that the data base could be enhanced, can think a little bit about the kinds of questions or the kinds of policies that we would like to know more about and be able to link in with the data, so that we can at least do some relational analysis. Not a question.

DR. FLORES: All the state assessment systems are about to go through peer review. So, do you anticipate maybe having to put an asterisk in the data base, PPR, post-peer review, because of changes that might be brought about to assessment systems?

PARTICIPANT: That is an interesting idea, Geno. I think what I would anticipate is that kind of an

important part of that peer review is going to be the evidence that states provide on alignment.

It strikes me that that would be useful, in some ways, to get information into the data base about the adequacy of alignment.

Now, it is alignment, admittedly, to different state standards. So, it is not alignment with the same thing, but at least it will tell you something about -- hopefully it will tell you something about -- how each state has done in having assessments that really reinforce whatever their state standards are. If there is a lot of variation in that, it would be nice to know about it.

PARTICIPANT: In general, I think that is one of the difficulties we certainly encountered using these data, was that we didn't always know when something about the testing system changed.

We would hear from somebody, "Oh, did you know that this state moved its cut score, and that is not in the data base anywhere?"

It is a little hard to think about how you would exhaustively put all of that into the data base, what form it would take.

I think that is something that, without that information, it is impossible to make any interpretations of what is going on. So, somehow that does need to be

incorporated either directly into the data base or as separate sort of descriptive documents that go along with it for each state.

PARTICIPANT: Has there been any systematic communication between peer review teams?

PARTICIPANT: I think the peer review teams themselves are independent of each other. Certain individuals serve on multiple peer review teams, not always in the same mix.

PARTICIPANT: What you really need is to combine grants coming up with the right questions, so that everybody raises the same points about each of the systems.

That way we might be able to move toward some similarities that would be an improvement. Some overlap would mean that those states will share some stuff that this overlapping guy or woman has.

A simple way would be to put up a listserv with access to the people who are appointed to peer review teams, and have them put up some of the things they are thinking about, and kick it around.

DR. SCHAFER: As a frequent peer reviewer, I think probably need to comment on this point. Yes, you are absolutely right, that there is a lot of shuffling that is going on among the peer review teams.

The peer reviews -- and the teams are fully aware

of this -- are advisory to the U.S. Department of Education. As such, there is a great deal of difference, I would guess, among the points that are made.

They try to be -- at least in my experience -- as comprehensive as possible, so that the Department of Education, then, will try to standardize the reviews. Now, that is not a perfect system, but there is that element in the process.

PARTICIPANT: I have about 24 points listed, but many of them have been made. So, I am only going to give six points that are kind of clarifications from where I see people are.

The first one is related to the idea of scale. We can always caution our measurements down to cut levels, but if we have, a priori, restrictive variance in the dependent measure, we necessarily restrict covariance. That is my message about scales. Everybody said the other things.

With regard to the matching, one of the things we don't necessarily have when we do a single match is some idea of how much non-treated matching groups would vary.

We know there would be some and, therefore, it is not the match group, when you are using existing data. Obviously, when you are gathering raw data, that can be very different.

When you are using existing data, multiple groups

or matches, at least those can be randomized into a few different matching groups, to get an idea of the degree to which matching groups might vary among themselves.

The comparison group, a lot of times I don't believe it is really appropriate to talk about the [rest of comment not caught by the microphone]. If you can, when you have existing data, get other comparisons that are maybe in the same district, or very, very similar, but a good distance away, so it is not a one or other type of thing.

There is also this idea of the definitive study. Someone mentioned that. With tobacco as a cause of cancer and heart disease, it is true that we did not have human experiments. However, we did have a great deal of in vitro experiments, where DNA was damaged, was inflected in cells, and in vivo experiments where exposure to tobacco smoke and components of it did cause, in an experimental manner.

My point is, a lot of times no study really does stands by itself, but adds to addition. When we can put it together with a meta-analysis, that is great, but we do still have other levels of things.

We might need to find out whether we think library stuff gets kids to read, at least to take books out, take them home, and that might go together with a different set of studies which come at a very different

level.

I also think that we have to be very careful to remember that there is the total error or uncertainty around a thing.

We have all the uncertainty. Some of it we can know directions of, and when we know directions of it, we have to be careful not to treat that as noise, but treat that as things we figure pretty much balances in the direction.

The last point that I want to relate is that relational studies can be very useful. Without some form of relation, there really can't be a cause.

There has to be some form of dose response or linear, or anything, there has to be a relationship, and correlational studies, and various forms of pattern detection studies can give us information in which to put the context of other studies.

So, a data system like this maybe can't give us the definitive causal answer, but it can give us some studies that can go into a set about which we make the rhetorical points about what kind of conclusions we are trying to draw. That was just my soap box responses to what everyone has been saying today.

DR. DUNBAR: Other comments or questions from the floor? All right, I am going to give you a homework

assignment, then.

I have been trying to put together a few notes based on today's conversation that seem to be points that have been made repeatedly, and maybe represent the kinds of, if not formal conclusions, certainly consensus discussion that has taken place.

I am just going to throw some of these out to get your reactions to them. Your homework assignment, which you can complete now, in the time we have left, so that you don't have to take it home with you, or you can work on tonight so that we can collect some of this information tomorrow, is to add to this list things that you heard today which you thought were consensus kinds of comments.

Number one, multiple reporting metrics, percents above cuts, percents within achievement levels, and some kind of standard scores are desirable.

Number two, systematic documentation definitions of relative covariates is critical. Tell me how broadly we need to define relevant, I think, would be the unstated question there.

Number three, quantitative analyses bolstered by careful observational methods are also important, because they will help us understand details of program implementation that otherwise would lead possibly to misinterpretations of results.

Number four, understanding what can and cannot be done with individual matched records helps to frame what we can and cannot do with school level data. This is the missing levels issue that Judy was talking about.

Number five, it would be very helpful to have in this data base information about within school variation. Those are my five points.

I nearly added a sixth about the resolve to create matched data sets at lower levels of aggregation, but nobody seemed to jump at that. So, I didn't think we were at any kind of consensus. You could work on that one.

DR. MCLAUGHLIN: A point on your within school variation. States aren't normally reporting that. So, you won't find much of it on the data base. I think there might be one or two states that have it for one or two years.

PARTICIPANT: They have it?

DR. MCLAUGHLIN: States have it, and we have been taking from the public publications. In fact, most of the states have multiple standards.

I have found that, in fact, you can get an approximate answer to the right question about variance by looking at the distribution of percentages.

PARTICIPANT: The data base, is it consisting exclusively of data that has been submitted by the state, or is whoever is keeping the data base keeping other

publicly available information, and is there a finite set of data, or is this going to be a variable data base as it matures?

DR. MCLAUGHLIN: What we have done is just to add the CCD data to it, figuring that most data sets for federal programs and many other programs have the NCES school code.

So, other things can then be merged with it. For instance, we have merged with NAEP, we have merged with the Schools and Staffing Survey, we have merged with -- I guess that is on school NAEP -- with the CSRD [Comprehensive School Reform] files.

Other contractors have merged with the charter schools and so on, but the idea was that we had that at least for this component of the data set, but then others could merge in aspects, characteristics of the schools.

So, no, except for CCD, we haven't added in other variables so far. I don't think that was in our plans, but I am hearing some things that would make -- as a contractor, AIR has to worry about what the client would buy and, since NAEP is paying for it right now, what we are doing, which is separate from EDEN, we have to make some rationale for why it would be important for comparing NAEP and state assessment results.

DR. DUNBAR: Just one follow up to the comment

about creating matched data sets at lower levels of aggregation, there is kind of a natural experiment going on in that right now in many states, at the state level.

Newly instituted systems of student IDs have been created. Geno was describing one to you. We have one in Iowa.

In fact, they are anonymous, in that kids don't know what their ID numbers are. So, there are elaborate administrative mechanisms in place to get those onto student records, but those are what are being matched.

The extent to which those kinds of data can be made publicly available, I think, is one of the questions that we might consider, and I am looking over at Diana Pullin now to help guide us through the ethics of that tomorrow in the synthesis discussion.

Any other questions or comments for today? Well, let's give all the presenters another round of applause for a great round of presentations, and thank you all for sticking with us for today.

We will hedge our bets for the weather tomorrow, and we are going to ask Stuart to tell us what the story is on that.

DR. ELLIOTT: The word from our hosts is that this building is going to follow the federal government rules. So, if nothing happens tomorrow, we will be here

tomorrow and we will begin at 8:30.

If there is some sort of a delay, which would be announced on the usual sorts of media -- for example, a two-hour delay -- then we will meet here, we will meet at a two-hour delay. So, we would start at 10:30. We will figure out how to adjust the schedule over the course of the day.

In the unlikely event that the federal government closes -- and this does seem like it is quite unlikely, I don't think that much snow is supposed to come, but if that happens -- this building will be closed.

The National Academies is loathe, once we have had a bunch of people coming to town for an event, to not have things go on.

So, in that event, we will try to find another location to meet. My hunch would be that we would be able to meet at the building we have, which is 500 Fifth Street, Northwest.

We will be able to find meeting space there, if the federal government is closed, because all of our other activities will also be sort of in slight suspension there.

That is close to Chinatown, so we would at least be able to find some place, even in a fairly reasonable blizzard, that is still open, because I have had the occasion to do that.

That building is also close to the subway. So, for those of you who are in town and staying at the Foggy Bottom hotel, it will probably be possible to get there.

In the event that the federal government closes, I will leave a message on my voice mail that lets you know what we are going to be doing. I hope to see you all here tomorrow morning.

[Whereupon, the meeting was recessed, to reconvene the following day, Friday, December 9, 2005.]