

**RESEARCH IN EDUCATION:
ON THE LEADING EDGE OF SCHOOL IMPROVEMENT?**

**A PUBLIC POLICY FORUM PRESENTED BY
NATIONAL EDUCATION KNOWLEDGE INDUSTRY ASSOCIATION
PROGRESSIVE POLICY INSTITUTE
EDUCATION QUALITY INSTITUTE**

**WITH SUPPORT FROM
ACADEMY FOR EDUCATIONAL DEVELOPMENT
THE KNOWLEDGEWORKS FOUNDATION**

TUESDAY, MARCH 26, 2002

**ACADEMY FOR EDUCATIONAL DEVELOPMENT
1825 CONNECTICUT AVENUE, NW, 8TH FLOOR
WASHINGTON, D.C.**

9:30 – 10:45 AM

"WHAT IS QUALITY, AND HOW GOOD IS 'GOOD ENOUGH?'"

**RICHARD WHITMIRE
USA TODAY**

**DAVID MYERS
MATHEMATICA POLICY RESEARCH**

**STEVE ROSS
UNIVERSITY OF MEMPHIS**

*Transcript by:
Federal News Service
Washington, D.C.*

JIM KOHLMOOS: I'm going to ask the panel now, the first panel, to move right up here. Just to let you know as they're moving up here, we will have a question and answer period in about a half-hour, towards the end of this panel. Lisa, I don't know if you are going to be around or not or do you need –

LISA TOWNE: I'm not leaving.

MR. KOHLMOOS: So if you have questions, be sure and hold them, write them down, so that we can engage in a very nice conversation, okay. And let's move right into the first panel. And also folks, there are refreshments in the back, and the restrooms on the side here. We won't view it as impolite if you need to go someplace temporarily.

Again, rather than go into lengthy introductions right here, right now, I'm pleased to introduce Richard Whitmire, who's going to moderate the panel discussion today. He comes from USA Today. He's done a marvelous job in raising educational issues on the editorial page for all of us, and so Richard, I present it to you.

RICHARD WHITMIRE: Well, I think Andy Rotherham invited me to do this not because what I know about education research, but because of what I don't know. If somebody like David Myers starts using terms like suggestive research evidence, then I'm supposed to roll my eyes and raise my hand and say, David, please speak in English. Until the last year or so, I wrote very little about education research which is the norm for most education reporters, but it's also kind of bizarre because most medical reporters I know deal with medical research all the time. I think it's an indication of how education research is perceived. In all honesty, it is perceived as second rate, even among reporters. But obviously this is changing; it's changing fast. You know a revolution is in the air when Congress uses the term scientifically-based research more than a hundred times in the ESEA reauthorization.

It's happened with bilingual education. This is a revolution brewing from outside the education system, which means it can be powerful but also produce some unintended consequences. But if it turns out right, it should bring education research up to the level of social science research such as that as we've seen with welfare reform.

My education in education research began a couple of years ago when I was – essentially started out to do a pup piece on some of the Obey-Porter whole school reform models. And I looked at the list at what schools were choosing and found that on the top five were some whole school reform models that I had never seen before. And I chose one of those and called the developer and said, please send me the research behind it, and a couple days later a nine-pound box arrived. And I started shifting through this and chose the research this developer thought was most cogent because it had been listed in the brochures, and I looked at this research, and even as an amateur, I could see that there were problems. The gains in three schools that they were talking about when compared with other schools, the gains were there but those same three schools were also using Success for All extended school day, after school, and extended school year.

In another case, I found a couple of teachers who had agreed to compare their classrooms in different schools, and a year later the teacher who came out better was working for the developer. So it was a wakeup call for me, and I started looking at other education issues and found that there was, in my opinion, no real research behind bilingual programs that had been running, no real search behind reading programs, teacher professional development programs, and the list goes on. It's not that I'd just become grumpy about bad research – editorial writers are supposed to be grumpy about things – but I had become a convert to what high quality research can do. I visited the First in the World Consortium Schools in Chicago when they were well underway there, and what they did was draw on the TIMS research and changed completely the way that they did math and science education. For example, they stopped repeating math items early in elementary school, and the science work there was just based on problem solving. Paul Kimmelman, who oversaw those successful reforms, is in the audience today.

So over the past couple years, there's been enough renewed interest in this topic to ask more sophisticated questions such as, how good is good enough, which is what we're supposed to answer today. And you kind of have to think about who's on the receiving end of that question. Take charter schools, for example. A parent wants to know one thing about a charter school – will this reform work for me – and let's say a legislator wants to know on a broad scale, is this a policy worth pursuing. Among all consumers of education research, I would argue that the principals are probably the most important consumers. People assume that the superintendents are making decisions, but in fact, when I tour schools what's happened is that the authority has devolved to the principal level as part of the whole accountability movement.

And so you have people who really were chosen years ago based on their skills of making the busses run on time and the cafeterias work choosing whole scale reform models. So I would argue that they are probably the least experienced to make this decision, and at least based on the early years of the Obey-Porter reform models, it seemed that too many principals were making these decisions based on who had the shiniest models at the regional fairs. So we haven't given that principal much to work with, but it's not as though they can put everything on hold until the perfect research is completed. They have to make the decisions immediately, which is the question for this panel, which is how good is good enough.

We have two people to answer that question. First is David Myers who's known to reporters for his research in voucher outcomes, which needless to say is second-guessed by everyone. Clearly David Myers is not someone who has learned to avoid controversy in his life. He's a senior fellow at Mathematical Policy Research, and associate director, and he's a nationally recognized expert in evaluation of education programs. Currently he's directing a ten-year evaluation of Upward Bound program.

And also Steve Ross, who is known to reporters as also someone who has not learned to avoid controversy because he invested many years in evaluating the programs in Memphis. As many of you know, many of those programs were terminated last year

by a new superintendent coming in -- this is the Whole School Reform experiments that were ran in Memphis -- but his research stands as incredibly valuable. So

[Technical difficulties.]

DAVID MYERS: But, really, I think the bigger question that most people want to answer is, will these schools compete for each other, and what happens to the kids who do use vouchers and those who don't use the vouchers. Charging to stay behind her, they can't find a school to go into. All of these experiments that you've seen reported in the last year or two, high quality work, they answer the former question, not the latter question. So how can I say vouchers work? Yeah, they work for African American kids in New York City who come forward and want to use it. We don't know the answer to the other question.

So I think the first thing you really got to answer yourself is, is it answering the relevant question for you? I think if you go a step below that, you go past the question and say to yourself, can the researcher convince me that they did a good job of comparing similar kids. Now you often say, and I think this is where the suggestive part gets in trouble -- you often say well, I'll compare participants against non-participants. That should be a red flag right away. There's a reason that other group didn't participate. They weren't motivated, they didn't have access, whatever. But you've got to say, how does a researcher make them comparable. Well, the simplest way, the one with the least assumptions: random assignment, use the experimental design.

All right, finally, how closely linked is the experiment to what we would expect to see in a full-scale implementation. I've got to the voucher example again. Here the vouchers are very small: \$1,400 a year typically, in the three experiments that have been going on. Yeah. Is that fit what we would expect to see in a full-scale public policy? Probably not. I'm running out of time? Okay. Too bad. I have the mouse.

All right, let me go ahead here. Last slide. You asked me to put this in, so... How to foster higher quality research? Again, in the beginning I said, I think that if we set the bar too low, people will say, whoa, you can't do random assignment. It's too hard, nobody will agree to it. That's the problem: nobody will agree to it, okay. I think if there's political backbone out there that you can get people to do it. There are a lot of evaluations that go in school settings. For example, we're doing evaluation of the Teacher for America program, randomly assigning kids to classrooms within schools, elementary schools.

So I think you can do it. But what you need to do is make the consumers -- teachers, principals, superintendents, and policy makers -- have a mindset that's similar to that of a researcher. They need to think like researchers. They have to have some doubt in what they're doing, and I think maybe Richard or Lisa said earlier, researchers are always skeptics. You need to have the consumers be skeptics, and I think one of the worst places you see this -- I'm really going to pick on people now -- is in schools of education. They have such strong ideology about what's right and what's wrong. They

don't distrust what they're telling teachers to do in the classroom. I often find when I go out if I can turn a teacher into somewhat of a researcher and say, do you really know what you're doing is making a difference for these kids. Could you make a bigger change in their life if you did something different? As soon as you can get a teacher into that position, and they don't feel threatened that they're failures, they'll often go with the rigorous evaluation. Thanks.

STEVE ROSS: Okay, I think we're ready. I've been told I have 30 seconds for this presentation, so it'll be done in a fairly compressed speech. I want to thank David for the striptease joke: you seem like a much better audience as a result, more receptive. I'd like to give you greetings from Memphis, former home of Elvis, former home of school reform, present home of the Mike Tyson fight where our image goes up every day.

(Laughter.)

Imagine if the only information physicians had about drugs came from the pharmaceutical companies, or from highly technical scientific journals. Imagine if the only information we had about cars came from the manufacturers or from highly technical scientific journals. Imagine. That's the way things often are in education, and I'd like to address some of those issues in talking about what works, and maybe where we can go. I've been told I can't go backwards, so I need to be sure I don't go too far forward, so concluding thoughts –

(Laughter.)

Thank you very much.

(Laughter.)

This doesn't work. Talk amongst yourselves. Okay, try the next one. Okay, go ahead. Okay, we are here. Some assumptions about evidence, and if you remember from about five minutes ago, I was talking about the gap between getting evidence to consumers. Some assumptions about evidence, and these comments are very consistent with what David talked about. We disagree on some minor issues, but we're very much on the same page with major things.

Evidence is very difficult for policy makers to interpret: more on that later. Studies differ in quality: more on that later. They differ immensely in quality, and do policy makers have the time and expertise to decide which are the best in quality. Data analysis and results are often complex. It is very difficult to read articles and differentiate between manova and mancova, and effect sizes and probabilities, and regression equations. I'm the editor of a leading educational journal. I use sophisticated reviewers with PhDs in research. Twenty-five percent of a time I get back one review of a study that says, accept pretty much as is, and another reviewing reading the same study that says, reject, it's not a good quality study. These are the experts. We're in a complex world in reading these reports.

Results are often mixed. In a typical study, ten schools go up using the model, but five go down. A very quick Memphis story, for every model that was used -- there was eight of them back in the early days -- there was at least one school that raised achievement significantly, and another school a mile away with the same kids, using the same model, where achievement went down. So what are we looking for? There are no simple answers. Results are often mixed, and the spin that is put on results often ends up being policy.

Results are often context specific. What works in Memphis may not work in Detroit, and may not work in Washington, may not work with a different population. Results are often not third party. The world of research is pretty incestuous, and it's hard for policy makers to differentiate between which researchers were working for the design team, and which were totally independent. Actually, we have very little true third party research for reasons I can go into later.

And now, for a bar graph, for your entertainment -- (laughter) -- but we won't spend very much time on this. Just to show you the different spins on the same results can result in totally different policy implications. This graph is from Memphis, and the commercial appeal used this graph to indicate that the designs were not working, CSR was not effective -- Comprehensive School Reform. This graph shows the mean percentiles of the schools, and I'm not going to bother to interpret this. If anybody's interested, I have copies and can send you it, but this design, this graph shows that Memphis did not improve medium percentile rank of its schools. And I can give you a lot of reasons why that's very hard to do.

Six months before, the commercial appeal used this graph, which is the exact same schools, but this shows improvement, gain scores, value added scores, to say that the designs were working. The same newspaper, the same schools, the same community, two totally different opinions in six months time, and the nation goes on those sound bites, not on reading the original research.

Suggestions: expert third party review seems the best root. And I don't want to be abstract. I need to leave you in the little time I have with the sound bites so that we can move ahead, maybe today. What I mean by expert third party review is the type of this of OELI is doing with its clearinghouse, the type of thing that EQI, Educational Quality Institute, is doing with its review. You need qualified expert reviewers to look at the research as a whole, and come up with an informed opinion of where it works. We can't have, in my opinion, policy makers and practitioners doing it: it's too complex, it's too confusing. Media and policy makers need to be educated on how to interpret and report evidence, and I hope we're starting that today.

And now for a slide that probably will be quite frightening. It's really quite easy -- you have a multiple-choice test on this. No, just kidding, but I do have copies of this. I don't expect you to be able to read this -- I really don't -- but this is an example of a rubric similar to what EQI will be using. This isn't a rubric that -- or something that says, count

the number of studies that produce positive results. This asks informed people to look at the evidence overall, and go through different dimensions and scales to determine where there is clear evidence, and where there isn't, and to come out with information for consumers because policy makers are challenged to do that on their own. If anybody would like copies of this, I brought about 40 with me, and I can easily email it to you.

Concluding thoughts: having qualified, expert, third party reviews like the clearing house and EQI will be doing, should drive program and model development and quality control. We don't have that today. What we have today is what Richard referred to, the nine pound box – or was it 90? The big box of data that the vendor sends. We need better information and a better process: we don't have it. Impetus for higher quality research should be provided by having those kinds of reviews. Evidence will be more consumer oriented; it is not consumer oriented today. Selections of programs and models will be more data driven: it is not data driven today. My center is currently working with about 500 CSRD schools. At many of them, the selection of models is not data driven; it is for other reasons. Probably more on that later.

Let me give you what I think are unanswerable questions, but it is what the media asks all the time, it's what the public asks all the time. In my opinion, if we go this route, we will not be leaving No Child Behind. Which program model is best is not a good question. Which program model is not a good question. How effective is program or model X, in my opinion, is not a good question. I think much better questions are, which program model can be implemented best in this context, and seems most effective for these desired outcomes. Do you want technology to be used? Pick Connect, and I'm not making an advertisement; I'm just saying where Connect seems to work. Do you want reading to improve for at risk? Look at Success for All, but don't ask which is best. We don't have an answer to that, and never will. How effectively can program, model X be implemented, and what are its likely effects in this particular context. I think when we ask these questions, we have information that consumers can use to leave no child behind. Thank you.

MR. WHITMIRE: In my time keeping we have 11 minutes left, so I'm just going to skip the summary and just ask two questions of the presenters, just a question a piece, and then we'll open it up for a couple of questions from the audience.

On thing I'm interested in with David Myers, who clearly is tired of the question, do vouchers work, and of course I'm going to walk right into this trap and say, do vouchers work. If I could hit on point two that he made, what do we need to know – what does a policy maker need to know to answer that question. What kind of studies have to be done still and of what type of study, for a policy maker to answer that?

MR. MYERS: Sure. I mean, when I think about what I would do next if I were a policy maker or a researcher in terms of trying to sort out the voucher question, I mean yea, I think – I hate to keep going back to suggestive because it suggests things now that I said that, but I would probably design a study where the unit of analysis I'm looking at is quite different. The school district, and I would have school districts where they are

saturated with vouchers and I would also have a control group without. I would use the experimental design there, and I think there is an opportunity to do it. I wouldn't do it on such a large scale – Tom Glennan's (sp) laughing at me: it's all right. I wouldn't do it on such a large scale, it'd be infeasible. I mean, it would be an incredibly expensive experiment. But the work that we see now that talks about competition and things, it's based on so many assumptions. Carolyn Hawksby's (sp) work, for example. She uses rivers and lakes and things like that to distinguish a place that could have high concentrations of private schools – it's really strange. It makes an interesting story, but I sure wouldn't want to make policy on it. But I would set up a full-scale evaluation with an experimental design.

MR. WHITMIRE: And it's doable, to answer the competition question.

MR. MYERS: I think it is, yeah.

MR. WHITMIRE: And the question I had for Steve, I really wanted to draw on his Memphis experience and take it back to the principal's perspective. And I really do think in most cases it is the principal trying to answer this. You mentioned here the right and wrong questions to be asking about whole-school reform model. Put yourself in the place of a principal, and different reform models are being shopped around: should I take Accelerated Schools, should I take Success for All. If you could take us through the red flags. As this principal is looking at the research, what are some red flags for that principal to watch out for, and what are some green lights?

MR. ROSS: Well, in 30 seconds or less, the red flags are data that comes from the vendor. Every vendor, every design team has positive data, every single one, that can show positive results where there was improvement. The green lights are evidence from trustworthy sources such as journals and such as true third parties, and success stories from people, from other principals, other schools, that are similar to yours, that you know about. I have tremendous respect for principals; I think it's one of the hardest jobs in the world, but for many principals, doing a reform means being told by the district that you have 30 days or less to write a proposal for the federal money or state money, and they find out very quickly what's hot, what's going on, put together a proposal, and wallah. They have a model that may not work three years down the road because it wasn't researched reflectively enough.

MR. WHITMIRE: We do have time for a few questions. If you could identify yourself briefly, and because we only have time for a couple of questions, if we could make them very brief questions. I know there is a tendency to give speeches sometimes, but people who have questions for either of the researchers. Yes, sir.

Q: I'm Ron Hanson (sp), and I have a question about – sorry, Ron Hanson, REOI. I have a question about the analysis. You've talked about a district level change, you've talked about the importance of principals. A lot of the language in the legislation looks at student level achievement. Obviously there are a lot of problems when you're trying to find student level achievement, but the change, the intervention is happening at

the district, for instance. How do you deal with that sort of problem in getting a good quality evaluation of something?

MR. ROSS: You brought that up.

MR. MYERS: We've done some work in terms of thinking it's an old idea now, but on the Reading Excellence Act, where schools were involved in it for example, but you're trying to affect student's reading achievement. And literally what it comes down to is collecting information on students. But in terms of an experimental design, you think about assigning the schools to the treatment and control conditions. So in a sense you're having data at multiple levels, and then trying to chart out the trajectories, the learning curves for these kids in terms of their reading. It's not that different, it's just that you're assigning the school to a condition, but you're looking at the kids' outcome. And there's some statistical issues, but –

MR. ROSS: Yeah, I was hoping David would say that because I didn't want to disagree, but I agree totally with what he said. If you don't look at student level data, it's very difficult to show any kind of effects on achievement, because what mostly accounts for achievement is student aptitude. What next accounts for achievement is student socio-economic status. What next accounts for achievement is teacher effectiveness. By the time all that variance is eaten up, there's very little left for programs. Unless we do precise analysis, we cannot see that little that's left, and bad studies are done all the time that conclude no significant differences for statistical validity reasons, invalidity, not because of that's a true finding.

MR. WHITMIRE: Yes, sir.

Q: (Off mike.)

MR. MYERS: Yes, and no. The evidence is a bit mixed on this, but what you find is – for example, there've been some comparisons of what goes on at the Tennessee Star on the school side research, and they weren't able to replicate the random assignment results using these other approaches. Sometimes you can get it to work; sometimes you can't. The problem is that we don't have a good understanding of when they work well and when they don't work well. So you're in this uncomfortable situation of not knowing if you got the right answer or not. Now, that's not to say that random assignment gives you certainty. All it does is it minimizes the chances that you are making a mistake. I mean by chance you could start off with two groups that are different, and it does happen sometimes, but you're in a much better position, starting off with the experimental design.

So we don't know enough about it. And you look through the literature and there have been meta-analysis, Will Shaddish (sp) at Memphis and some other people, and they say that on average, if you do it enough, you may come up with the same answer. The problem is, we don't replicate things much in education. So on any one instance, you could be way off the mark.

MR. WHITMIRE: Yes, ma'am.

Q: -- This comment is directed toward Steve Ross's last comment which is that it is not sensible to ask the simple question, does this program work, or is this program effective, but rather he posed a more complicated question: does this program work on behalf of certain outcomes for certain kids. But I wonder if you think, given the right research, you can sensibly ask and get an answer to this simple question: does it not work?

MR. ROSS: That's a simple question? (Laughter.) Yes, through the kind of process which David's talking about which we don't have patience for in education, and that's to do the replication. If a particular design -- an examination of a program is replicated under many contexts, and if there was more time we could talk about educational innovations like the new math that did not work in replication after replication. So yeah, it'll be possible to eliminate if we have time and a process like I was talking about with EQI and the clearing house where experts can look at that results because believe me, even for a non-working program, the vendor, as I said before, will have plenty of evidence to tell America that it works.

MR. WHITMIRE: Yes.

Q: Larry White with Houghton Mifflin Company. Is there the capacity for third party research, and if there isn't now, how long would it take to gear up, to be able to do all the third party research that No Child Left Behind would call for?

MR. MYERS: I don't think there is the capacity. I've talked to people, for example, at the Hewlett Foundation, and they're interested in looking at assessing in a rigorous way, the impacts of assessing various curriculums, and they even worry that there isn't the capacity out there to do it. You look at the research firms who really -- I mean, there's the university side, and then there are the research firms such as the Mathematicas, Rand, Apt, and others, and I don't think we have the capacity to do it at this point. It'll take years to develop. The same researchers are out there, churning the waters that have been there for the last 15, 20 years. There's not that much of an influx of new people, and I think it's a problem.

MR. WHITMIRE: Yes, sir.

Q: Charles Rankin, Kansas State University. We've been using the term, scientifically based research and evidence based education, and effective program evaluation. Do you see these terms as interchangeable, and if not, what are the differences and what are the implications for how to approach them?

MR. ROSS: The terms probably aren't totally interchangeable, and I'd have to go through a more semantical analysis than we have time for, and I hope this at all addresses your question. In my opinion, it's very hard to define scientifically based research and

evidence based decision-making. A lot of the public thinks that it's black or white, that a study is valid or invalid, and it's much more complicated than that. What I feel generically is that scientifically valid research is – and I don't want to throw jargon at you – but is research that has internal validity. In other words, the effects can be attributed to programs and not to extraneous factors and random assignment helps. But today, unlike 30 years ago, valid research has external validity, meaning it can play in New York and it can play in Peoria, although in different ways. But I'm sure we can get into a more detailed discussion, and you're after something that, I'm sorry, I can't address enough in the time we have given the very subtle differences.

MR. WHITMIRE: I think we have time for one more question.

Q: -- You know we often talk about what would be nice to do in education research without dwelling nearly enough on what we have the capacity to do. Larry Snow White asked the question about human capacity. I would ask the question about financial and political capacity, in the context of David Myers suggestion of what it would take to do proper voucher research. To do the kind of voucher experiment that you were suggesting, you would need a maybe \$5,000 per kid over a number of years in a sizeable place, and you would need a public funder to come up with that kind of money because no private funder could do it, and you would need to political willingness of districts to be compared with each other on a random basis, which means that districts would have to volunteer for this thing, and be chosen and be included or not included. It strikes me that you've just suggested something that we don't have the political capacity to do either of those two things, either to find the public funder that wants to do that kind of large scale, long term voucher research and can get approval to do so, nor the districts that are willing to volunteer in that way. So is this not a naïve and impractical suggestion you are making, David?

(Laughter.)

MR. ROSS: Yeah, I think so.

MR. MYERS: I think we're in a time checker.

(Laughter.)

No, I wouldn't say it's naïve. I agree with everything that you're saying. It is particularly an issue of political will, to invest in this kind of research, and I think that goes across education. It almost always comes down to the political will to do something. I think the financial side of it is – I mean it would be huge, but if you can get past the political side of it, of getting people to commit to do this, and to do this, I think the financial side would be somewhat secondary to it. My concern with education research is that we often go in and say, it's too hard to do the right thing, and I often will draw a line in the sand because I want the bar to be high enough. And I think there will be situations where you have to argue against certain kinds of designs, but we often fall too far away from it, and we do something that's not good. Suggestive: it's the strip

tease. So I would try to argue that – and I would try to work with states, for example, to set it up within states, for example. So, no, I don't think it's naïve: hard, but not naïve.

MR. WHITMIRE: Well, Checker Finn always gets the last question, so I think we'll dismiss the panel and thank you very much for the great presentation.

MR. KOHLMOOS: Richard, Steve, David, thank you so very much for taking time out of your busy day to be with us. Now just a second, a little choreography here. We're going to take a – here's a controversial thing – a five-minute break. So you can stand up. If you really need to excuse yourself please do, but we're going to come back in five minutes, and the next panel can make their way up now.

(END OF 9:30 – 10:45 AM PANEL.)