1-1-2005

The 2004 Claremont Debate: Lipsey vs. Scriven

Stewart I. Donaldson
Claremont Graduate University

Christina A. Christie
Claremont Graduate University

The 2004 Claremont Debate: Lipsey vs. Scriven

Determining Causality in Program Evaluation and Applied Research: Should Experimental Evidence Be the Gold Standard?

Stewart I. Donaldson and Christina A. Christie

Claremont Graduate University

While there is little disagreement about the need for, and value of, program evaluation, there remain major disagreements in the field about best practices (Donaldson & Lipsey, in press). For example, Donaldson and Scriven (2003) invited a diverse group of evaluators to Claremont in 2001 to share their visions for "how we should practice evaluation" in the new millennium. Theorists and practitioners discussed a wide range of views and evaluation approaches, many at odds with one another, on how best to improve evaluation practice (e.g., the experimental paradigm, evaluation as a transdiscipline, results-oriented management, empowerment evaluation, fourth generation evaluation, inclusive evaluation, theory-driven evaluation and the like). In response to some of the heated exchanges, Mark (2003) noted "it seems ironic when evaluators who espouse inclusion, empowerment, and participation would like to exclude, disempower, and see no participation by evaluators who hold different views." He

further concluded that whatever peace has been achieved in the so-call quantitative-qualitative paradigm wars remains an uneasy peace.

This uneasy peace seemed to revert back to overt conflict in late 2003, when the U.S. Department of Education's Institute of Education Sciences declared a rather wholesale commitment to privileging experimental and some types of quasi-experimental designs over other methods in evaluation funding competitions. At the 2003 Annual Meeting of the American Evaluation Association (AEA), prominent evaluators discussed this new level of support for experimental designs as a move back to the "Dark Ages" of evaluation. Subsequently, the leadership of the AEA (supported by Michael Scriven among many others) developed a policy statement opposing these efforts to privilege randomized control trials in education evaluation funding competitions:

AEA STATEMENT

November 24, 2003

> Dear Colleagues,
>
> We encourage AEA members to share their views on Scientifically Based Evaluation Methods with the U.S. Department of Education. Up to now a number of members have shared their views with other members on EvalTalk. This discussion has been helpful in clarifying our thoughts and in presenting potential arguments, but NOW it is time for AEA members to share their views directly with the Department of Education.
>
> A statement has been prepared by a team of distinguished evaluators including: Randall Davies, Ernest House, Cheri Levenson, Linda Mabry (chair), Sandra Mathison and Michael Scriven. This team received valuable assistance from: Lois-ellin Datta, Burt Perrin, Katherine Ryan and Bob Williams. We are grateful

to this team for their rapid response to this proposal. This statement has been approved by the current and future Executive Committees of the Board of the American Evaluation Association, including:

Molly Engle, 2002 President

Richard Krueger, 2003 President

Nick Smith, 2004 President

Sharon Rallis, 2005 President

Nanette Keiser, 2002-2003 Treasurer

Kathleen Bolland, 2004 Treasurer

We encourage AEA members to share their thoughts directly to the U.S. Department of Education and possibly with legislative leaders. If you agree with the AEA statement, you might indicate your support of the AEA statement.

OR

If you wish to offer other arguments or points of views, please submit those as well.

Responses are to be sent to:

Margo K. Anderson, U.S. Department of Education, 400 Maryland Avenue, SW., Room 4W333, Washington, DC 20202-5910

Or by internet to: comments@ed.gov and include the term ``Evaluation'' in the subject line of your electronic message. Comments must be received on or before December 4th.

Sincerely

Richard Krueger, President

American Evaluation Association

* * * * *

American Evaluation Association Response

To U. S. Department of Education

Notice of proposed priority, Federal Register RIN 1890-ZA00, November 4, 2003

"Scientifically Based Evaluation Methods"

The American Evaluation Association applauds the effort to promote high quality in the U.S. Secretary of Education's proposed priority for evaluating educational programs using scientifically based methods. We, too, have worked to encourage competent practice through our Guiding Principles for Evaluators (1994), Standards for Program Evaluation (1994), professional training, and annual conferences. However, we believe the proposed priority manifests fundamental misunderstandings about (1) the types of studies capable of determining causality, (2) the methods capable of achieving scientific rigor, and (3) the types of studies that support policy and program decisions. We would like to help avoid the political, ethical, and financial disaster that could well attend implementation of the proposed priority.

(1) Studies capable of determining causality. Randomized control group trials (RCTs) are not the only studies capable of generating understandings of causality. In medicine, causality has been conclusively shown in some instances without RCTs, for example, in linking smoking to lung cancer and infested rats to bubonic plague. The secretary's proposal would elevate experimental over quasi-experimental, observational, single-subject, and other designs which are sometimes more feasible and equally valid.

RCTs are not always best for determining causality and can be misleading. RCTs examine a limited number of isolated factors that are neither limited nor isolated in natural settings. The complex nature of causality and the multitude of actual influences on outcomes render RCTs less capable of discovering causality than

designs sensitive to local culture and conditions and open to unanticipated causal factors.

RCTs should sometimes be ruled out for reasons of ethics. For example, assigning experimental subjects to educationally inferior or medically unproven treatments, or denying control group subjects access to important instructional opportunities or critical medical intervention, is not ethically acceptable even when RCT results might be enlightening. Such studies would not be approved by Institutional Review Boards overseeing the protection of human subjects in accordance with federal statute.

In some cases, data sources are insufficient for RCTs. Pilot, experimental, and exploratory education, health, and social programs are often small enough in scale to preclude use of RCTs as an evaluation methodology, however important it may be to examine causality prior to wider implementation.

(2) Methods capable of demonstrating scientific rigor. For at least a decade, evaluators publicly debated whether newer inquiry methods were sufficiently rigorous. This issue was settled long ago. Actual practice and many published examples demonstrate that alternative and mixed methods are rigorous and scientific. To discourage a repertoire of methods would force evaluators backward. We strongly disagree that the methodological "benefits of the proposed priority justify the costs."

(3) Studies capable of supporting appropriate policy and program decisions. We also strongly disagree that "this regulatory action does not unduly interfere with State, local, and tribal governments in the exercise of their governmental functions." As provision and support of programs are governmental functions so, too, is determining program effectiveness. Sound policy decisions benefit from data illustrating not only causality but also conditionality. Fettering evaluators with unnecessary and unreasonable constraints would deny information needed by policy-makers.

> While we agree with the intent of ensuring that federally sponsored programs be "evaluated using scientifically based research . . . to determine the effectiveness of a project intervention," we do not agree that "evaluation methods using an experimental design are best for determining project effectiveness." We believe that the constraints in the proposed priority would deny use of other needed, proven, and scientifically credible evaluation methods, resulting in fruitless expenditures on some large contracts while leaving other public programs unevaluated entirely. Statement prepared by: Randall Davies, Ernest House, Cheri Levenson, Linda Mabry (chair), Sandra Mathison and Michael Scriven. This team received valuable assistance from: Lois-ellin Datta, Burt Perrin, Katherine Ryan, and Bob Williams.

Opposition to the AEA Statement

An influential group of senior members of the American Evaluation Association opposed the AEA Statement, and did not feel they were appropriately consulted as active, long-term members of AEA. In response to President Krueger's call for members to share their individual views on this matter, a new statement now referred to as the "NOT AEA STATEMENT" (as seen on Evaltalk) was submitted to the U. S. Department of Education:

NOT THE AEA STATEMENT

Posted on Evaltalk on: 12-3-2003

> AEA members:
>
> The statement below has been sent to the Department of Education in response to its proposal that "scientifically based evaluation methods" for assessing the effectiveness of educational interventions be defined as randomized experiments when they are feasible and as quasi-experimental or single-subject designs when they are not.

This statement is intended to support the Department's definition and associated preference for the use of such designs for outcome evaluation when they are applicable. It is also intended to provide a counterpoint to the statement submitted by the AEA leadership as the Association's position on this matter. The generalized opposition to use of experimental and quasi-experimental methods evinced in the AEA statement is unjustified, speciously argued, and represents neither the methodological norms in the evaluation field nor the views of the large segment of the AEA membership with significant experience conducting experimental and quasi-experimental evaluations of program effects.

We encourage all AEA members to communicate their views on this matter to the Department of Education and invite you to endorse the statement below in that communication if it is more representative of your views than the official AEA statement. [Comments can be sent to the Dept of Ed through Dec. 4 at comments@ed.gov with "Evaluation" in the subject line of the message].

****

This statement is in response to the Secretary's request for comment on the proposed priority on Scientifically Based Evaluation Methods. We offer the following observations in support of this priority.

The proposed priority identifies random assignment experimental designs as the methodological standard for what constitutes scientifically based evaluation methods for determining whether an intervention produces meaningful effects on students, teachers, parents, and others. The priority also recognizes that there are cases when random assignment is not feasible and, in such cases, identifies quasi-experimental designs and single-subject designs as alternatives that may be justified by the circumstances of particular evaluations.

This interpretation of what constitutes scientifically based evaluation strategies for assessing program effects is consistent with the presentations in the major textbooks in evaluation and with widely recognized methodological standards in

the social and medical sciences. Randomized controlled trials have been essential to understanding what works, what does not work, and what is harmful among interventions in many other areas of public policy including health and medicine, mental health, criminal justice, employment, and welfare. Furthermore, attempts to draw conclusions about intervention effects based on nonrandomized trials have often led to misleading results in these fields and there is no reason to expect this to be untrue in the social and education fields. This is demonstrated, for example, by the results of randomized trials of facilitated communication for autistic children and prison visits for juvenile offenders, which reversed the conclusions of nonexperimental studies of these interventions.

Randomized trials in the social sector are more frequent and feasible than many critics acknowledge and their number is increasing. The Campbell Collaboration of Social, Psychological, Educational, and Criminological Trials Register includes nearly 13,000 such trials, and the development of this register is still in its youth.

At the same time, we recognize that randomized trials are not feasible or ethical at times. In such circumstances, quasi-experimental or other designs may be appropriate alternatives, as the proposed priority allows. However, it has been possible to configure practical and ethical experimental designs in such complex and sensitive areas of study as pregnancy prevention programs, police handling of domestic violence, and prevention of substance abuse. It is similarly possible to design randomized trials or strong quasi-experiments to be ethical and feasible for many educational programs. In such cases, we believe the Secretary's proposed priority gives proper guidance for attaining high methodological standards and we believe the nation's children deserve to have educational programs of demonstrated effectiveness as determined by the most scientifically credible methods available.

The individuals who have signed below in support of this statement are current or former members of the American Evaluation Association (AEA). Included among us are individuals who have been closely associated with that organization since

its inception and who have served as AEA presidents, Board members, and journal editors. We wish to make clear that the statement submitted by AEA in response to this proposed priority does not represent our views and we regret that a statement representing the organization was proffered without prior review and comment by its members. We believe that the proposed priority will dramatically increase the amount of valid information for guiding the improvement of education throughout the nation. We appreciate the opportunity to comment on a matter of this importance and support the Department's initiative.

Signed by:

Leonard Bickman

Professor of Psychology, Psychiatry, and Public Policy at Vanderbilt University, Associate Dean, and Director of The Center for Mental Health Policy at the Vanderbilt Institute for Public Policy Studies; Coeditor of the Sage Publications *Applied Social Research Methods Series* and the *Handbook of Applied Research Methods* and the editor of the Journal, *Mental Health Services Research*; recipient of the American Psychological Association's Public Interest Award for Distinguished Contribution to Research in Public Policy and the American Evaluation Association Outstanding Evaluation award; past president of the American Evaluation Association.

Robert F. Boruch

Professor in the Graduate School of Education, Fels Institute for Government, and the Statistics Department of the Wharton School of Business at the University of Pennsylvania; Fellow of the American Statistical Association and the American Academy of Arts and Sciences; recipient of the American Evaluation Association Myrdal Award for Evaluation Practice and the Policy Studies Organization's Donald T. Campbell Award; founder of the Evaluation Research Society, a parent to the current American Evaluation Association.

Thomas D. Cook

Joan and Serepta Harrison Chair in Ethics and Justice and Professor of Sociology, Psychology, Education and Social Policy at Northwestern University; Coauthor of *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, and *Foundations of Program Evaluation: Theories of Practice*; Fellow of the American Academy of Arts and Sciences and the American Academy of Political and Social Science; recipient of the American Evaluation Association Myrdal Award for Evaluation Science, the Donald Campbell Award for Innovative Methodology from the Policy Sciences Organization, and the Distinguished Scientist Award of Division 5 of the American Psychological Association.

David S. Cordray

Professor of Public Policy and Psychology at Vanderbilt University; Coauthor, Evaluation methods for social intervention, Annual Review of Psychology; past President and Board Member of the American Evaluation Association.

Gary Henry

Professor of Public Administration and Urban Studies, Political Science and Educational Policy Studies at the Andrew Young School of Policy Studies, Georgia State University; Coauthor of *Evaluation: An Integrated Framework for Understanding, Guiding, and Improving Policies and Programs*; former Editor-in-chief of *New Directions for Evaluation*; recipient of the American Evaluation Association Outstanding Evaluation award and the American Society for Public Administration and Center for Accountability and Performance Joseph Wholey Distinguished Scholarship Award; Board Member of the American Evaluation Association.

Mark W. Lipsey

Director of the Center for Evaluation Research and Methodology and Senior Research Associate at the Vanderbilt Institute for Public Policy Studies; Coauthor of *Evaluation: A Systematic Approach*; former Editor in Chief of *New Directions*

for Program Evaluation; recipient of the American Evaluation Association Lazarsfeld Award for Evaluation Theory.

Peter H. Rossi

Stuart A. Rice Professor of Sociology and Professor Emeritus at the University of Massachusetts at Amherst; Fellow of the American Academy of Arts and Sciences and the American Association for the Advancement of Science; Coauthor of *Evaluation: A Systematic Approach*, *Thinking About Program Evaluation*, and *Program Evaluation in Education, When? How? To What Ends*?; recipient of the American Sociological Association Commonwealth Award and the American Evaluation Association Myrdal Science Award.

Lee Sechrest

Professor Emeritus of Psychology at the University of Arizona and founder of the Evaluation Group for Analysis of Data; recipient of the American Evaluation Association Myrdal Award for Evaluation Practice and the Distinguished Scientific Contribution Award from the Division of Evaluation, Measurement, and Statistics, of the American Psychological Association; past president of the American Evaluation Association and the Division of Evaluation, Measurement, and Statistics of the American Psychological Association.

****

The 2004 Claremont Debate

The exchange above about the role of randomized control trials in program evaluation practice in educational settings set the stage for the 2004 Claremont Debate.

The apparent resurgence of issues reminiscent of the well-known quantitative-qualitative paradigm wars in evaluation has the potential to be destructive and to

stunt the healthy development of the discipline and profession. In an effort to seek a deeper understanding of the current dispute, and to possibly discover a middle ground or productive resolution, Claremont Graduate University hosted a debate between representatives from both sides. Below, you will find selected excerpts from the opening remarks by Mark W. Lipsey (who plans to publish a more complete version of his thoughts in the near future), followed by excerpts from the response from Michael Scriven.

Selected Excerpts from Mark Lipsey's Opening Comments

"In this context, it seems to me that there are at least three topics that we might discuss."

"One has to do with the way randomized trials appear in government agencies and the legislation and so on, some of which is simplistic and inept, as uncharacteristic as that is of government activity."

"Another thing we might talk about is the little flack in the American Evaluation Association (AEA) that involves the stance that was taken last year opposing an obscure division of the Department of Education to try to bring in some randomized evaluations to some of the projects it was funding. Since this event is being sponsored by an AEA Affiliate, that is a possibility. I'd be happy to explain to you why I think the AEA now has the same relationship to the Field of Evaluation as the Flat Earth Society has to the Field of Geology."

"The third thing we might talk about is the methodological issue and what is actually at stake in these methodological critiques. That is actually what I want to talk about, but if anyone, maybe the audience, or Michael wants to talk about the others, then I'd be happy to do that."

"We really are poorly served by this gold standard terminology. I think that when you use randomized experiments, which I am basically going to defend in this context, they are much like what Winston Churchill once said about democracy. He said, 'It's the worst form of government except for all the others that have been tried from time to time.' I do not think this is the gold standard. I think that for impact assessment randomized experiments are the worst methodology except for some of the others that have been tried from time to time. That is pretty much my theme here."

"Experimental and quasi-experimental designs have been around a long time and have well known properties. What's really new is this broadside against them from certain research communities."

"This issue has evoked mostly a yawn in areas where intervention research and program evaluation is done broadly. So, in mental health, public health, drug prevention, medicine, chronic delinquency evaluations, and a whole range of areas this is not a particularly exciting topic where randomized field trials are well respected, well known, widely used, and understood to be something of the state of the art for doing impact assessments. The reactions I've seen have come predominantly from the education research culture and to a certain extent from one wing of economists that work in this field that have an interesting take on it. I will get to that later on."

"Let me turn now to the non-experimental approaches. This is an area that has fascinated me. Back when flap was going on, methodological pluralism was all over the Evaltalk. I kept asking respondents and finally gave up on what these other methods were that were supposed to be equally valid, and the most interesting list came out: epidemiological methods, observational correlation

modeling, realist methods, case studies, qualitative, ethnographic, Glasser and Strauss' grounded theory, and from Michael Scriven the modus operandi technique, forensic analysis, direct observation, all put forth in establishing the effects of programs."

"I have in recent years, every time I see somebody putting forward the argument that qualitative methods could be used to assess program effects, I've been writing them for some examples. Show me a case where this was done convincingly."

"Why is the education research culture so riled up about randomized experiments? Here are a couple of possibilities. In all the politics this year, the Bush Administration, the Department of Education, the No Child Left Behind Act, there's a lot not to like there, okay? They have been pushing for randomized designs, so we may as well not like those too. The biggest factor I think is ideological. The education research culture bought into constructivism and post modernist epistemologies and so on really big time and there is a lot of ideological opposition. Tom Cook calls it science phobia to quantitative methods and experimentation and so on. Third, I think that there is a considerable amount of ignorance, not stupidity, not stupidity, but ignorance."

Selected Excerpts from Michael Scriven's Response

"Well, apart from the character assassination at the end, which I can tell you in the education community there may be people in it about which those things can be said, but the greatest attacks on constructivism are from people within the education community. So, there are plenty of others like us who absolutely reject all of that crap and so, it is certainly not true. Some of my friends are also on the side of the angels over there, like Tom Cook, for the new move. So, no, I don't think that is really a very plausible account of the story."

"I think that if you want to look at reasons why people objected, the three big ones are these. One, the objections were not at all against randomized control trials (RCT), they were against the decision to take all $500 million dollars of their research money and pull it out of anything except randomized control trials. Now, it is quite clear the previous speaker is not identifying himself with this extreme wing, but who is the leader of the extreme wing? It is the guy who is the head of the Institute of Educational Science that has the $500 million, and what does he say? He says there is no scientific way of establishing causation except by randomized and allocated control group trials, etc. etc. There is no such thing as scientific research in the area of human behavior except by means of RCTs, and that is complete bullshit! It happens to be coming from the guy who has all of the money. So, the sad thing is that this is man killing off alternatives"

"Read Tom Cook on problems in practice of running RCTs. So, this is a very tricky procedure. While it has theoretical advantages, the theoretical advantages in validity aspects of it are undeniable. That is not the issue. The issue is not whether or not there is an alternative that has the same theoretical bulletproof-ness. The question is whether there is an alternative that can get you results beyond reasonable doubt, and that is another story all together. Very often, you can get results beyond reasonable doubt in other ways."

"First, the concessions. We have not used RCTs when we should have many, many times. There have been many occasions when we could have pulled off RCTs, when we could have staffed them with competent people, and this is still the case in the present, and that was the best design around. The arguments around are sloppy arguments including a number of arguments that Professor Lipsey ran into at the Evaltalk discussion. There was a lot of whistling in the dark going on there and ideological crap going on. You have to get down to the logic of the cases and

you can't just pull this off by waving things like constructivism, observational, or etc. So, this is a situation where there is no doubt at all. This is a very powerful tool, and sometimes much the best tool, but it has as the same value as the torque wrench in a good mechanic's toolbox. For certain tasks, you can't beat it. After all, this is a quantitative instrument. The torque wrench reads out in inches and meters and so on, so this is very important if you are interested in matching the specs that you are supposed to be matching…a very good instrument. Nothing can match it, but it has a very narrow range of uses. Now, that doesn't matter if the alternative approaches aren't very good, but of course there is a lot of them and some of them are very good indeed."

"Well, there's a lot more I'd like to say, but perhaps I can just leave it by saying I think I agree strongly with him. A lot of the attacks have been empty and they have lacked specific examples that will work. A lot of the attacks are based on ideological positions, which are logically unsound. All of this is true, but nevertheless, given the difficulties facing RCTs, one has to be very cautious going to any sort of wholesale commitment to them. I hope in the future we can develop a better kind of existence than what we have at the moment."

Conclusion

Somewhat surprisingly, Lipsey and Scriven agreed that randomized control trials (RCTs) are the best method currently available for assessing program impact (causal effects of a program), and that determining program impact is a main requirement of contemporary program evaluation. However, Scriven argued that there are very few situations where RCTs can be successfully implemented in educational program evaluation, and that there are now good alternative designs for determining program effects. Lipsey disagreed and remained very skeptical of

Scriven's claim that sound alternative methods exist for determining program effects, and challenged Scriven to provide specific examples. Streaming video of the entire Claremont Debate can be viewed at: http://www.cgu.edu/pages/465.asp.

About the Authors

Stewart I. Donaldson is Dean and Professor of Psychology, School of Behavioral and Organizational Sciences, at Claremont Graduate University. He has published widely in evaluation, developed one of the largest university-based evaluation degree, certificate, and professional development programs, and has conducted evaluations for more than 100 organizations during the past decade.

Christina A. Christie is an Assistant Professor, Director of the Masters of Arts Program in Psychology and Evaluation, and Associate Director of the Institute of Organizational and Program Evaluation Research in the School of Behavioral and Organizational Sciences at Claremont Graduate University. Her research interests focus on investigating the relationship between evaluation theory and practice and issues related to the development of descriptive theories of evaluation. She has conducted a variety of educational evaluations, and evaluations of social programs targeting high-risk and underrepresented populations.

References

Donaldson, S. I., & Lipsey, M. W. (in press). Roles for theory in contemporary evaluation practice: Developing practical knowledge. In I. Shaw, J. Greene, & M. Mark (Eds.), *Handbook of evaluation*. London: Sage.

Donaldson, S. I., & Scriven, M. (2003). Diverse visions for evaluation in the new millennium: Should we integrate or embrace diversity? In S. I. Donaldson &

M. Scriven (Eds.), Evaluating social programs and problems: Visions for the new millennium (pp. 3-16). Mahwah, NJ: Erlbaum.

Guiding Principles for Evaluators (1994). New Directions for Program Evaluation (No.66). San Francisco: Jossey-Bass.

Joint Committee on Standards for Education Evaluation (1994). The program evaluation standards: How to assess evaluations of educational programs. Thousand Oaks, CA: Sage.

Mark, M. M. (2003). Toward a integrative view of the theory and practice of program and policy evaluation. In S. I. Donaldson & M. Scriven (Eds.) Evaluating social programs and problems: Visions for the new millennium (pp. 183-204). Mahwah, NJ: Erlbaum.