jmde
Journal of MultiDisciplinary Evaluation

http://evaluation.wmich.edu/jmde/ *Articles*

Why an Active Comparison Group Makes a Difference and What to Do About It

Lois-ellin Datta
Datta Analysis

The Randomized Control Trials (RCT) design and its quasi-experimental kissing cousin, the Comparison Group Trials (CGT), are golden to some and not even silver to others. At the center of the affection, at the vortex of the discomfort, are beliefs about what it takes to establish causality. These designs are considered primarily when the purpose of the evaluation is establishing whether there are outcomes associated with a program and, if so, how confidently the results can be attributed to the program. If one concludes these designs are superior to alternatives for establishing causality, and have no more bad habits than the alternatives, then the RCT and the CGT are the methods of choice.

Much has been written about the advantages of the RCT and the CGT with regard to issues such as ethics, feasibility, and inference (Boruch, 2000; Cook, 2006). Ethical issues supporting use of RCTs include the injustice of continuing ineffective treatments for service recipients who might be better helped with other approaches. Feasibility issues include evidence from thousands of RCTs conducted in a wide array of circumstances showing that these designs can be carried out in the real world and are not limited to laboratory settings (see, e.g., the archives of the Cochrane and Campbell groups and many articles in *Evaluation Review*). Inference issues include the assertion that, other things being equal, the RCT is the best way to rule out biases that could under-estimate or over-estimate true effects of the treatment. There is a growing body of head-to-head comparisons that are consistent with these claims, particularly with alternative quantitative designs.

Much also has been written about the limitations of the RCT and the CGT with regard to issues such as ethics, feasibility, and inference. Ethical issues have included whether, if the approach being tried out is likely to work, it is right to deny service to some but not others. Feasibility issues have included whether enough participants will agree to randomization to develop an adequate sample and questions of whether, if examined carefully, the studies in the Cochrane and Campbell archives would prove to have been carried out in ways consistent (for example) with the American Evaluation Association Standards. Issues of inference have included whether knowing something does or doesn't work tells enough about what is happening inside that belabored image, the black box, and whether, if tested head-to-head against other alternatives proposed such as observational techniques (Scriven, 2006), the RCT would indeed prove better in controlling for biases or even just as good.

One does not have to look far for masterful analyses of the concerns with RCTs. Indeed, some of the most trenchant come from those well-versed in the quantitative, experimental approaches. Lipsey and Cordray (2000), for example, have elegantly discussed variables beyond experimental control. They write,

Random assignment…is recognized as a

> useful means of equating groups prior to delivery of an intervention… [However] whereas assignment to treatment and control conditions is a defining event in outcome evaluation, decades of experience have shown that after assignment important processes occur that can seriously influence the quality of the evaluation design, the interpretability of the results, and the utility of the study (p. 346).

Among other concerns, they discuss poor program implementation, augmentation of the control group with non-program services, poor retention of participants, receipt of incomplete program services by participants, attrition, and "…a host of participant characteristics…[that] can interact with exposure and response to treatment in ways that further complicate the situation" (p. 346).

I focus here on one after-assignment condition that may notably affect the logic of the RCT and the CGT designs, particularly the central assumption that, all other things being equal, observed differences if any between experimental (E, treatment) and non-experimental (C, control, comparison) groups are attributable to the treatment. My concern might be characterized as augmentation of the control and experimental groups with relevant non-program services in non-random, potentially biasing, ways. Somewhat more attention will be given to the experiences of the C group because of this group's particular significance for the logic of the RCT.

The Three Main Points:

1. In human service programs, the C groups are likely to be active, rather than passive. Ditto the E groups. This seems particularly likely if the RCTs are zero-blinded as they very often may be in human service areas such as education, health, and welfare.
2. It matters if the groups are active, because this can lead to non-random augmentation of services particularly for the Cs but also the Es. If so, then charmed by the apparent rigor of the RCT and CGT designs, we risk invalid conclusions being proclaimed with near-undaunted certainty in the executive summary and text, with some caveats. The real stuff seems to appear in footnotes and appendices. (See, for example, Ludwig & Phillips, 2007, pp. 20-21.) In RCT theory, it is not expected that life is confined to a petri dish. However, for the logic to work, E and C groups should not have similar experiences, and the other post-assignment factors affecting the E and C groups must be random, not biased toward one or the other group. The presence of other relevant factors may add variability, but design integrity is maintained *if the additional variability is randomly distributed.* Design integrity is threatened if non-random post-assignment factors add greater variability to one group or to another. Design integrity, by definition, is lost if E and C experiences converge. One would predict, in a meta-analysis, an inverse relation between effect size and treatment convergence.
3. Since the best assumption for human service programs may be an active C group, the evaluator, like Hamlet, should take arms against this sea of troubles both prospectively and retrospectively, in ways I will describe. The discussion will assume that taking arms does not necessarily mean doing something else with regard to design, although this should be considered, but possibly adapting the RCT and CGT designs to make them more useful. Donaldson and Christie (2005) observe:

> 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 (p. 64).

With regard to vulnerability associated with active E and C groups, ideally, there would be considerable empirical information about sound alternative methods, looked at with the same rigor of logic and experience as has been applied to debates about the RCT and CGT. Apart from Yin's and Bamberger's fine work, there seem to be relatively few publicly available reports of qualitative or mixed method large-scale evaluations where the purpose was attribution. With some exceptions, such as Roger's masterful discussion of Appreciative Inquiry, few alternative model outcome evaluations, even at the yellow-polka-dot bikini scale, have been honored with such detailed analyses as House has bestowed on the RCT/ CGT evaluations of Sesame Street, Follow Through, and Jesse Jackson's programs. Having a Scriven-Fetterman-Greene Collaboration acquire 1,300 or more examples of alternative design evaluations to be examined in such depth would possibly be a grand step forward.

Happily, more is known, meta-evaluation wise, about such quantitative alternatives as regression discontinuity and interrupted time-series designs, particularly through the work of Shadish. The regression discontinuity design increasingly may be regarded as comparable to the RTC in internal validity, and in establishing attribution.

What Are We? Chopped Liver?

"What are we? Chopped liver?' can be translated approximately as "I am not a potted palm" and "We don't get no respect." In evaluation, the Chopped Liver Effect could mean accepting the notion of a passive C and E group, who will largely stay in place, thus maintaining treatment differentiation as required by the RCT. Yet both C and E groups may not be potted palms, but rather may be actively engaged in determining what happens to them.

There are at least four ways in which the differentiation in experience required by the logic of the RCT can be compromised: (1) Es do not receive intended treatment, (2) Cs receive the intended E treatment, (3) Cs receive treatments very similar to the E treatments, and (4) both Es and Cs receive, in non-random ways, other or additional treatments similar to the intended E experiences. When both non-receipt of intended treatment for the Es and receipt of E treatment by the Cs occur, the situation has been described as cross-contamination.

It happens.

In 1968, the Children's Television Workshop, building on the popularity of shows such as Captain Kangaroo and the ubiquity of television sets, set its sights on promoting school readiness. The show was named *Sesame Street* and has since become one of the most widely seen of all children's programs.

Then, the world was not so sure. Joan Ganz Cooney and her colleagues garnered support from the Corporation for Public Broadcasting, the U.S. Department of Education, Project Head Start, and the Carnegie Corporation. The funders wanted formative evaluation to help develop the show and summative evaluation to see whether it worked. I thought that green frog was mighty cute but questioned whether the children would learn much. A RCT design was possible, with the Corporation for Public Broadcasting making the first year of Sesame Street available only in selected communities. The evaluation, carried out by Educational Testing Service under Sam Ball, showed the children indeed learned, with those from middle and upper income families learning more, possibly because the parents were hooked, and watched the show along with the children,

reinforcing its lessons.

Even then, the control families were not entirely passive. Where they could, the families watched Sesame Street. Some went to considerable lengths to get their children in preschools which received the broadcasts. The extent to which this blurred the effects of the show was difficult to estimate (Bogatz and Ball, 1971).

The active C group for Miss Piggy and Kermit is not unique.

1. The Abt evaluation of the Comer program in Detroit found, at the end of the implementation period, almost total overlap in the extent to which Comer principles were carried out between the E and C schools. In some C schools, principals and teachers decided if this was good for the E schools, it was good enough for them. In other schools, the reforms initiated district-wide by the Detroit school system part-way through the experiment reflected many of Comer's ideas (Millsap et al., 2000). Similar observations, also for a randomized experiment testing Comer effectiveness, were made by Cook and his colleagues.
2. Orwin et al. (1994) and Lipsey and Cordray (2000) have documented the active control group effect in their massive true randomized design test of treatments for homeless men with multiple problems. The men were homeless, not stupid, despite drug, alcohol, and mental health problems. The Cs figured out which treatments were being offered where, and how to get enrolled in the ones they preferred. Ditto some of the Es. In the end, the carefully constructed, meticulously sampled, years-in-design awesomely costly experiment did not have enough "true" Es and Cs for the planned data analysis.
3. The even more massive Congressionally mandated national randomized control test of whether Head Start works also involves a carefully constructed, meticulously sampled, years-in-design experiment (GAO, 2003; ACF, 2005). The design involved random assignment to Head Start or non-Head Start conditions, pre and post testing, and then assessment after first and third grades: in essence, the Westinghouse-Ohio design but "fixed" to control for initial differences. About 4,000 applicant families, selected after meticulous sampling, were told in 2002 their three and four year olds could enter the Head Start program to which they applied (N = 2,500) or "Sorry, they can't attend" (N = 1,900). A few months after the random assignment to E and C conditions, a survey showed about 50% of the C children were enrolled in other programs, and 18% (that's over 350 C children) had already enrolled in other Head Starts. And 14% of the E children did not use Head Start during the year 1 study period when the initial benefits of the program were to be established.
4. The High Scope/Weikart preschool programs tested in Follow Through surprisingly (to the data analysts) showed little or no evidence of benefits even on measures reported so glowingly in the initial studies carried out by the developer himself in his own pilot program (Schweinhart & Weikart, 1993). Further investigation showed that some comparison sites had, with the support of the developer who passionately believed his program helped children, adopted the High/Scope approach. Again, the residual Ns were too small for analysis.
5. And, in an interesting twist on the active group, Parker, Asencio, and Plechner (2006) found a carefully design intervention "failed" for a juvenile treatment because the program was so attractive to those assigned at random, they didn't want to "graduate" and deliberately sought recidivism to be re-assigned to the treatment, while the controls used their street smarts, too, to get into the program. After describing the consistent no-difference findings, the authors observe,

> …our data suggest they [the youth] did everything they could to have another chance at PREP [the intervention]. The fastest way to gain access to PREP was to fail at the placement. Going AWOL, being expelled from a placement, or even committing another crime once released from what was considered to be a successful placement, all of which the treatment group did at higher rates than the control group, constituted a path back for these youth into Juvenile Hall, where they would ask and plead to be sent back to PREP (p. 53).

6. Random assignment through the justice system to drug treatment court vs. usual sentencing might seem to offer little opportunity for E and C activism. Gottfredson et al. (2007) report, however, that in a carefully thought-through RCT study in Baltimore, about 9% of the Es did not receive treatment and about 7% of the Cs scheduled for usual sentencing received drug court treatment.

Probably as pervasive as such activism is the MCIYE effect: My Control Is Your Experimental. I am evaluating a National Science Foundation supported test of the value of bringing together career-track science graduate students and public school teachers. There are five schools, six teachers, and six Fellows involved in year 1, reaching over 200 students. In each of years 2 and 3, about 10 teacher/graduate fellow pairs and over 400 students yearly will participate. The graduate fellows, who receive an excellent financial assistance package, take an extra year of graduate school. During this year, in addition to their graduate studies, they are partnering with the teachers to adapt science curricula to local conditions and learn about science education in the schools. Will this benefit the teachers, fellows, and students?

An RCT or CGT, comparing participating and non-participating teachers might seem attractive. If one could not assign at random from equally willing volunteer teachers for a RCT, perhaps some teachers could be matched reasonably closely on what might appear to be relevant variables of a CGT. Would such an effort be likely to reduce biases and increase certainty?

First, what else is happening that is relevant and it is sufficiently random to permit some comparisons? An inventory of what else is happening suggests it will be hard enough to sort out why the participating classes look the way they do, let alone establish some meaningful comparisons. Some but not all of the teachers in the five schools and possible comparisons are part of the National Science Foundation EPScOR teacher training program. Some but not all schools are implementing an America's Choice curriculum that infuses science education with academic basics. Some but not all teachers are part of a Harvard Graduate School of Education teacher education and school reform project. Some but not all teachers are involved in special training efforts such as the summer on-the-water voyages. One school received a $1,000,000 private donation to improve education. Two other schools are elbow deep in infusing science education in school garden projects. Some schools hav e a science fair initiative program, and a privately funded science education initiative is going great guns in the area---and this is just for starters. By the end of the three year project, some of these may have fizzled out, some may become super-novas, some may be emerging, and none of this is likely to be the random background noise between the E and Comparison schools that makes "all things being equal." There are clumps and clusters of experiences, not all benign, that like leaves in a current, swirl and regroup in multiple, complex, unstable and biasing patterns.

This is discussed by Cook et al. as "contamination." It certainly *is* contaminating with regard to the RCT and CGT designs, but the active control group is more than that and, I believe, it does make a difference for analyses

and conclusions through at least two effects: (1) narrowing the difference between E and C conditions in experiences actually received and (2) increasing variance.

Table 1 summarizes some of the possible post-assignment events identified by Lipsey and Cordray, and their likely effects on variance. The table is hardly definitive but may be a step toward a systematic consideration of the effects of post-assignment events on various evaluation designs and eventually, an estimate in a meta-analysis of their frequency and magnitude.

Table 1
Some Post-Assignment Events and Their Likely Effects on Variability

Primarily Affects	Events	Likely Effects on Variability
Controls	Attrition	Depends
Controls	Treatment leakage	Increases
Controls	Augmentation of experience	Increases
Controls	Cross-over	Increases
Experimentals	Weak implementation	Increases
Experimentals	Boosted treatment from non-program staff	Increases
Experimentals	Attrition	Depends
Experimentals	Multiple non-program treatments	Increases
Experimentals	Cross-over	Increases

Why the Active Control Group Matters

Tests of the reliability of an observed difference basically compare the observed differences between two or more groups against the average variability within each group. By definition, whatever reduces variability will increase the likelihood that a given difference would be identified as "significant," rare if only chance were operating. Vice versa, whatever increases variability will decrease the likelihood that an observed difference of a given size is attributable to the treatment.

Consider the Head Start case. The context of what else is happening includes welfare reform. Welfare reform requires many parents of preschool children to go to work. For many, that involves child care. Assume that most of the children selected for Head Start will remain in the program although moving, preferences, and family situations will have their effect; assume that at least 18% of the Cs wind up in Head Start; assume that about 50% in all enter other preschools, day-care centers, or other non-parental care. Head Start strives for high standards of quality through a rigorous monitoring and program review system. Assume this works fairly well, so that while some Head Start programs may be better than others in terms of program quality, the range is not huge. In comparison, alternative child care programs can be expected to have considerably more variability in quality. Some may be located in states that strive to equal or excel l Head Start standards. Others may not. Program quality *as experienced by the C children* is likely to be more variable than program quality *as experienced by the E children.*

So what? The "so what" is that program quality has been shown to be related to child development. Better program? Better outcomes

for the children, not always but in general. QED: the active C group for the Head Start randomized experiment is likely to have higher variability in measured outcomes than the E group. This could lead to finding a possible macro-negative (no or a small reliable difference) effect of Head Start as an E treatment versus the non-Head Start C group as a whole. This would be particularly true if the analyses were conducted according to Intent to Treat (ITT) rather than according to effects of Treatment on the Treated (TOT) (See Ludwig & Phillips, 2007, pp. 3-4.) Thus, the risk of under-estimation of Head Start benefits.

The effect of actual experience on outcomes is hardly a new finding, whether identified as treatment frequency, treatment intensity, or treatment quality. For example, in 1975, Stallings and her colleagues showed that the macro-negative effect of Follow Through was due in part to the differential ease of properly implementing different curricula. The evaluators at Stanford Research Institute had good observational data on the extent to which each curriculum variation was carried out. When comparisons were made between high quality/well implemented classes versus lower quality/less well implemented classes, several note-worthy findings emerged. High quality trumps low quality, regardless of the curriculum used. When only high quality sites are compared, some curricula clearly did better than others. At that time, we weren't smart enough: no data were collected on the comparison classes.

Millsap and her colleagues (2000) in the Comer study referred to found similar results almost 20 years later. Comparison of Comer versus non-Comer schools for learner outcomes showed no reliable differences. When high quality, well-implemented classes versus lower quality, less well-implemented classes were compared, ignoring the ostensible labels, students in classes using Comer principles learned more than students in classes not using the Comer principles. When high quality classes only were compared, students in the Comer schools did better. If Millsap et al. had not conducted these analyses, then the conclusion would have been that the Comer principles were ineffective.

In the Gottfredson et al. study, analyses by intended treatment yielded a few statistically reliable results favoring the E group (drug court). Analyses by treatment actually received (1) dramatically decreased the likelihood the observed differences were due to chance, p. values going for example, from .306 to .007, and (2) shifted findings on the nine outcome indicators from differences significant at the .05 level on only two indicators to finding all nine significant at the .05 level or less.

Treatment actually received—not the labels or the intent-to-treat—seems to be what makes a difference. Comparing outcomes only on the basis of the intended treatment can make for unnecessary death or discouragement by evaluation. The logic of the RCT and CGT designs require unbiased estimates of variability, that whatever else is happening in program and policy space in addition to the treatment, is happening to equal degrees and with equal intensity to E and C groups, and is happening at random. Equating *without evidence* intended treatment with actually received treatment is not likely to assure this.

What To Do?

This is not a diatribe against RCT or CGTs. Far from it. I think the value of these designs, when attribution is wanted and when appropriately used, has been demonstrated more fully at this time than the value of alternative designs for large-scale, high-stake impact evaluations. This is particularly so when the RCT and CGT are used in conjunction with methods aimed at getting inside the black box, with methods intended to enrich understanding such as appreciative inquiry and case studies, and with approaches integrating knowledge of what else is happening and what may emerge that complex adaptive systems frameworks may offer. The designs also may be particularly

suited to double-blind conditions, to well-defined treatments, to fairly brief interventions, to situations where there are meaningful no-treatment circumstances, and to groups fairly likely to be passive, rather than active.

These points are hardly new. There is wide agreement on these among evaluators inclined to enthusiasm for the RCT (see, e.g., Cook, 2006; Cook & Payne, 2002; Shadish, Cook, & Campbell, 2002) and those who are not.

In human service programs, it seems to me excessively heroic to assume an unbiased estimate of variability achieved through randomization that is not subsequently biased by non-random, relevant additional factors in policy and program space. These factors can close the treatment-as-experienced gap and increase C group variability, particularly through an active C group and an active C group probably should be assumed in many instances.

What to do? In the Head Start study and others cited, evaluators have decided it is better to risk under-estimation than to risk over-estimation of program effects. This is not a position I find persuasive, although as in much else, it depends on such specific circumstances such as how the observed effect sizes, costs, and other benefits for one type of intervention compares with those from alternative policies and strategies. For example, how do effect sizes in measures of child development compare for early childhood education and those for employment and training programs, housing, and income support? What are the benefit/cost comparisons? Considering the noise in measurement, in implementation, in what else is happening, sometimes it can seem rather astonishing if any signal comes through.

Several approaches have been used. One is to analyze the data with and without no shows and cross-overs, reporting both findings, the approach taken by Gottfredson et al. Another, which is thoroughly discussed in the Head Start impact study planning and first-year reports, is to analyze only by intent to treat (ITT) while acknowledging the possibility of estimating the residual effects by reducing effect size in proportion to no-shows and cross-overs. Thus, E/(1-n-c) where E is effect observed, n is the no-show rate (no shows/total treatment group n) and c is the cross-over rate (cross-over controls/total control n). (See Angrist et al., 1996; Bloom, 1994).

> When the basic impact estimates are divided by this factor (1-n-c), this expands into estimates of the effect of treatment on the just treated…If the assumptions of "no effect" on the "no shows" and of identical impact on the "cross-overs" and "cross-over-like" members of the treatment group are correct, this adjusted estimate of the effect of the treatment on those treated is just as reliable as the original impact estimate for all assigned (AFC, 2005).

The Angrist et al. adjustment seems to depend on several assumptions regarding availability and quality of services essentially equivalent to Head Start, and is expected to under-estimate Head Start effects, or, in similar studies, other treatment or experimental effects. With a large impact, large Ns, and low no shows and cross-over percents, the approximation may be close enough for credibly reducing uncertainty and sound policy guidance. As the cross-overs increase and assumptions become more dubious, the correction cure may be worse than the cross-over disease as a guide to policy. It seems to me we can do better as evaluators. Table 2 shows an alternative approach.

Table 2
What to do when an Active C Group Seems Likely

Prospective questions	Get information	Incorporate into design choices
1. How serious is this threat?	Survey what else is happening in program and policy space. Determine if these threats are likely to work against the treatment or work with it. Are they are likely to be random between E and C or are they likely to be systematic and biasing?	If bias is likely, add this threat to the list of factors to be considered when design choices are made. Are there alternatives that would not be similarly dismayed?
2. How could the evaluator get reliable estimates of the treatments actually received?	Are there ways--surveys, observations, interviews--that feasibly and reliably could document the treatment actually experienced by E and C groups? Scriven recommends such interviews, particularly carried out by evaluators with qualitative skills.	If there are ways that can be identified, include the time and costs of collecting and analyzing these data in factors to consider when design choices are made. Are there alternative design choices that would not be similarly dismayed?
3. How and how well could analyses by adapted to determine effects of the treatments actually experienced?	What analyses can be planned to take into account information on treatment experienced by the C and E groups? How robust, statistically, are the methods for taking this information into account? What are the consequences for statistical power of needing to make these analytic adjustments?	The costs and feasibility of making analytic adjustments should be considered when design choices are made. Specific plans should be made for what analyses will be run, in what sequence, such as conducting exploratory analyses of the C group data looking for within C group patterns related to more treatment/less treatment received, if the Cs were active. If the costs are high and the analytic consequences unknown, try a pilot study and work these out before the larger study is undertaken or considered the next best design alternative.
Retrospective questions	Information gathered	Actions to be considered
1. Does the evidence show active control groups?	Check the data collected based on design phase decisions. If none collected, try interviewing service providers as close as possible to the action to see what they know.	If no evidence of such threats, no need to go through analytic contortions. These can just run unnecessary risks of finding 5% of the time statistically reliable results at the .05 level.
2. If there is evidence of active control groups, find out the effect of the treatment-as-experienced.	Ideally, the evaluator has reasonable information of the treatment actually experienced.	Carry out the analyses planned during the design phase to check the effects of the treatment-as-experienced by E and C groups. Look, for instance, for evidence of within C-group differences in outcomes associated with different levels of treatment-like experience; compare C and E groups with about the same levels of treatment actually experienced. If using "conservative" analyses such as intent-to-treat regardless of treatment as experienced, show upper and lower estimates of effect using more and less conservative assumptions.
3. No data?	Get best estimates of probable degree of control group activity.	Run hypothetical analyses showing the range of best and worst cases with regard to program effects.

This costs of this approach would be higher than the Angrist et al. and similar adjustments because reasonably fine-grain information would have to be obtained on the experiences of both E and C groups. It seems more defensible, however, where extensive no-shows or cross-contamination may be expected, or where , as Shadish (2000) has noted, "…the programs may have reached so many of the potential participants that outcome evaluations might be thwarted in finding appropriate control group participants if a controlled design is used."

The Case of the National Head Start Evaluation

The next report of this evaluation, presenting the children's prowess in the first grade, is expected in 2007. The Head Start site describing this study, including its history in the General Accounting Office report asserting randomized designs were necessary for conclusions about effectiveness and the subsequent Congressional mandate, laudably has extensive information on the design, the measures, and the analytic plans (ACF, 2005; SRCD, 2005).

According to the early reports, much data will be available on the program-as-experienced by the Head Start Es. Some data, at a macro descriptive level (the parent-reported categories of whether the child experienced center-based child-care, was cared for at home, etc.) will be available for the C children. The Urban Institute, through its outstanding national evaluation of the New Federalism legislation and welfare reform, has access to state information on child care alternatives, child care policies, and some observational data that might be brought to bear. The evaluators plan to take experiences into consideration to the extent they can. In all, the evaluation may represent the high water mark for an RCT in the context of a mature, widely available national program.

This is among the most high-stakes of the national impact evaluations currently underway, with an almost $7,000,000,000 program budget (that is seven billion annually) affecting about 900,000 children yearly: large enough to be greatly concerned if a truly less-than-effective program is continued or greatly concerned if a truly effective one is judged disappointing. Ludwig and Phillips (2007) have compared results using ITT and TOT estimates, finding that while ITT is what is discussed in the report and much public debate, the TOT analyses show larger, more consistent differences all favoring Head Start, and that the effect sizes are favorably cost/beneficial, leading to overly pessimistic interpretations. The statistical nuances and views on the robustness (or lack of it) of statistical adjustments with about 33% cross-over can appear in policy space as academic wrangling, and the pessimistic conclusions enter citations, references, beliefs, and actions. This is not trivial, however, for policy.

"Head Start fails" or "Modest Head Start Benefits" may mean re-definition still further of Head Start as an academic preparation program to be administered by the public schools. "Head Start works" may mean continuation of the vision of a comprehensive, developmental program involving health, nutrition, parent involvement, and a wide range of developmental areas. This affects not only the 900,000 children enrolled in Head Start but also, indirectly, the over 5,000,000 million children age 0 through 5 from low income families. While not served by Head Start, these children may be affected through federal day care standards and state standards for child care and early childhood development programs. That is because these standards are influenced by what Head Start, the flagship, establishes as its requirements. The debate has begun, in discussions such as those by the Society for

Research on Child Development (SRCD , 2005) and Besharov of the American Enterprise Institute (2005) of the year 1 findings. They see, respectively, the glass either as half-full or as half-empty.

In my view, the policy space regarding preschool programs for low-income children made the RCT design an inappropriate application to begin with. This was, however, a Congressional mandate and a possibly unhappy instance of legislative micro-management of evaluation design. It remains to be seen whether the hard-working evaluation team will be able to apply a generally satisfactory statistical fix or if a future article by a future Campbell and a future Erlebacher will be titled, "How Active Control Group Errors Mistakenly Made Head Start Look Ineffective."

References

Administration for Children and Families. (2005). *Head start impact study and follow-up.* Washington, DC: Administration for Children and Families. Retrieved January 21, 2007 from http://www.acf.hhs.gov/programs/opre/hs/impact_study/reports/imptstdy_interim/I

Angrist, J., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444-472.

Besharov, D. J. (2005). *Head Start's broken promise.* American Enterprise Institute for Public Policy Research. (October 25, 2005).

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review*, *8*, 225-246.

Bogatz, G. A., & Ball, S. (1971). *The second year of Sesame Street: A continuing evaluation.* Princeton, NJ: Educational Testing Service. (See also Sesame Street Summary. ERIC, 1972, ASIN: B0006W2LS6).

Cook, T. D. (2006). Describing what is special about the role of experiments in contemporary educational research: Putting the "gold standard" rhetoric into perspective. *Journal of MultiDisciplinary Evaluation*, *6*, 1- 10.

Cook, T. D., & Payne, M. R. (2002). Objecting to the objections to using random assignment in educational research. In F. Mosteller & R. F. Boruch (Eds.), *Evidence matters: Randomized trials in education research.* Washington, DC: Brookings Institution.

Department of Education. (2005). Scientifically based evaluation methods. *Federal Register*, *70*(15), 3586-3589.

Donaldson, S., & Christie, C. (2005). The 2004 Claremont debate: Lipsey vs. Scriven. *Journal of MultiDisciplinary Evaluation*, *3*, 60-66.

Gottfredson, D. C., et al. (2007) The Baltimore city drug treatment court: 3-year self-report outcome study. *Evaluation Review*, *29*, 42-64.

Lipsey, M.W., & Cordray, D. C. (2000). Evaluation methods for social intervention. *Annual Review of Psychology*, *51*, 345-375.

Ludwig, J., & Phillips, D. A. (2007). *The benefits and costs of Head Start* (Working Paper 12973, available at www.nber.org/papers/w12973). Cambridge, MA: National Bureau of Economic Research.

Miller, L. B., Dyer, J. L., Stevenson, H., & White, S. H. (1975). Four preschool programs: Their dimensions and effects. *Monographs of the Society for Research on Child Development.*

Millsap, M. A. et al., (2000). *Evaluation of the Detroit's Comer Schools and Families Initiative.* Cambridge, MA: Abt Associates.

Mosteller, F., & Boruch, R.F. (Eds.). (2002). *Evidence matters: Randomized trials in educational research.* Washington, D.C.: Brookings Institution.

Orwin, R., Cordray, D., & Huebner, R. N. (1994). Judicious application of randomized designs. In K. J. Conrad (Ed.), *New directions for evaluation: Critically evaluating the role of experiments* (vol. 63, pp. 73-86).

Parker, R. N., Asencio, E. K., & Plechner, D. (2006). How much of a good thing is too much? Explaining the failure of a well-

designed, well-executed intervention in juvenile hall for 'hard-to-place" delinquents. In R. N. Parker & C. Hudley (Eds.), *New directions for evaluation: Pitfalls and pratfalls: Null and negative findings in evaluating interventions* (vol. 110, pp. 45-57).

Schweinhart, L. J., & Weikart, D. P. (1993). *A summary of significant benefits: The High/Scope Perry Preschool study through age 27*. London, England: Hodder and Stoughton.

Scriven, M. (2007). *The logic of causal investigations*. In Press.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*.

Shadish, W. (2000). Evaluation theory is who we are (1998 Presidential Address). *American Journal of Evaluation*.

Society for Research in Child Development. (2005). *Placing the first-year findings of the national Head Start impact study in context*. Office for Policy and Communications.

Stallings, J., Almy, M., Resnick, L. B., & Leinhardt, G. (1975). Implementation and child effects of teaching practices in follow through classrooms. *Monographs of the society for research in child development*.

U.S. General Accountability Office. (2003). *Education and care: Head Start key among array of early childhood programs but national research on effectiveness not completed*. July 22, 2003, GAO-03-840T.