COMPARISON GROUPS AND SURVEY RESEARCH

Ken Watson[1]

This paper deals with the desirability of designing surveys in such a way that results can be compared to previous existing data. The writer explains why there are practical difficulties in assessing the significance of data collected in a one-time survey where these data stand alone and are not readily comparable to other existing data, i.e., where control group data or other benchmarks do not exist.

There is a story attributed to Anthony Downs about the horse and rabbit stew. There was a cook who set out to make this stew. He meticulously thought through and described the characteristics he wanted in his rabbit, carefully identified the population of eligible rabbits, and then selected at random his typical rabbit. The elegance of his rabbit selection was a thing of beauty. Unfortunately, the cook had little sympathy with horses, and taking any old horse, simply threw him in the stew. Much survey research is like the horse and rabbit stew - the part that is done well tends to be overwhelmed by the horse.

It seems to me that the horse in the stew is generally the lack of comparison groups. Of course, before we explore this idea further, it is necessary to point out that sometimes we are dealing with just plain rabbit stew: much survey research is simple counting. Tim Thompson has pointed out in his paper that approximately 25% of Statistics Canada's survey activity is in the area of general purpose statistics. I take this to mean the kind of simple counting that is typical of census activities. On the other hand, 37% of statistical activities were concerned with program planning, operation or policy accounts, and a further 31% of program activity was concerned with program evaluation. This is our horse and rabbit stew - survey research to support evaluation and strategic planning.

The horse and rabbit stew is not simple counting, although counting is involved. It is concerned rather with identifying the effects of some action. It is tempting to think that counting "effects" is just like counting noses, and that all survey research which we are dealing with is just plain rabbit stew. Fortunately or unfortunately, this is not so.

---

[1]President, Ottawa Public Policy Research Group Ltd.

Over the past year, I have participated in or observed closely 6 or 7 survey research projects undertaken by different agencies in the federal government. All of them were concerned with identifying cause and effect in some way. They were concerned not just with looking at the world to see how many tall people, or healthy people, or happy people there were; but rather to identify how many more tall, healthy, or happy people there were because of a given government policy or program. Only in one case could one distinguish the rabbit in the stew.

The survey researchers examined the participants in a government program and found that 70% of them were more than 10 pounds overweight. 45% of them thought that the program was a good thing, and 63% had mothers who were living. We find that the people in the stew are interesting in all kinds of ways, which may or may not be pertinent, but finally the question has to be asked, "so what?". Traditional survey research has identified, for example, the incomes of participants in a given program, but has not been able to answer the "so what?". The critical question of course is, what is the incremental or additional part of incomes attributable to the program with which we are concerned.

There is nothing mysterious about the process. It is simply a matter of having a basis of comparison. Without such a basis, the survey researcher is simply whistling in the dark. The interesting question then is, why year after year we continue with survey research designs which do not include decent comparison groups. I don't say 'perfect', just 'decent'. Most of the survey research projects which I have observed were without any basis of comparison whatsoever. In the remainder of this paper, I would like to consider why this may be so.

Technical Difficulty

Perhaps it is technically very difficult to get good comparison groups in survey research. This idea is sometimes reinforced by the jargon. The technical term for a piece of survey research with a good comparison group is a RCFTS, Randomized Controlled Field Trials. This sounds forbidding to say the least, and involves a concept of "randomization" or "random allocation to treatment and control groups" which has two built-in problems - the concept is continually confused with the concept of random sample, on one hand, and,

on the other hand, has connotations for the non-researcher that imply a mindless and perhaps absurd way of deciding who would receive the program and who not. So a rather opaque jargon which tends to promote misconceptions may be part of the problem.

But given that one hires professionals to design survey research projects, let us consider whether it is technically difficult for them to design research efforts which include decent comparison groups. Essentially the answer is no. In the final analysis, no pun intended, a good comparative design is far simpler than anything else. In fact, weaker non-comparative designs tend to become either trivial or extremely complex, because of their limitations. There are statistical techniques which reportedly "adjust out" differences between groups and which seek to equate groups which differ initially. The techniques, which essentially involve matching program participants and non-participants with respect to their demographic or other characteristics through co-variance or regression analysis, are sophisticated but require strong assumptions about the underlying nature of the data, and these assumptions are seldom valid.

Another possibility is that there are insurmountable practical difficulties in randomly allocating people to treatment (program) groups and control (comparison) groups. For example, one might not know how many applications there are going to be and applications might arrive in trickle fashion from diverse sources all over the country. There may be logistical problems in constructing comparison groups. The application approval procedure is often complex and time-consuming as it is, without introducing an additional complexity. Also, the procedures to establish basic eligibility may be quite extensive. Since we want only eligible persons in both the treatment and comparison groups, the administrative system might be strained logistically to generate enough eligibles to both absorb all of the program funds in the treatment group, and still have enough eligibles for a comparison group. But ultimately, this seems a little far-fetched. For one thing, one often has a certain flexibility in deciding upon what basis the randomization will be done. For example, some of the possibilities are randomizing by time period, by community, by project, or by participant. Identifying and constructing a good comparison group takes imagination, but given a decision on design, the logistics are seldom insurmountable.

There are some technical problems which are peculiar to comparison group designs. The principal ones are the problems of attrition and contamination. Constructing a comparison group is just the first and simplest step. Keeping it together over any period of time and keeping it truly separate from the treatment group is much more difficult. To consider one example, I was involved in the latest stages of the Manitoba Basic Annual Income Experiment. Initially, the comparison group in this project contained approximately 2,400 families. Over a period of three years, attrition due to family mobility and other factors had cut this number to approximately 1,400. With rates of attrition like this, comparison groups can dwindle away very fast. Nevertheless, this is a problem of longitudinal projects stretching over several years rather than a problem particularly related to comparison group designs. There is a de facto relationship because randomized field trials have tended to be longitudinal and multi-stage in their data collection. The second problem of contamination is a real one, but difficult to generalize about. Its dynamics are closely related to the nature of the program or treatment. Again, it requires a little imagination and planning. Given this, one can generally protect the integrity of a comparison group design.

The most important technical problem is the problem of "weak treatments". A little common sense will often tell us that a program "treatment" is so weak that it will be impossible by any means to isolate its effect from the influence of other much stronger factors outside of the program. Consider an imaginary case where a parole officer sees 95 ex-prisoners each month for one half hour each. The principal goal of this activity is to reduce recidivism. One might reasonably suppose that even if this program has positive results, they are likely to be so small as to be completely swamped by much stronger outside influences.

Of course if one is going to use survey research in examining the parole program, it is better to have a good design rather than a bad one. The problem is that even a good design is probably not going to be able to identify the facts because the treatment is relatively weak. In fact, in cases where good comparison group designs have been used to evaluate government programs, generally in the United States, the result has tended to be "no effect identified". There is a certain pernicious aspect to this. If the survey research design

is clearly a rigorous one, and no effect is identified, then there is a presumption that there is no effect in fact. Of course, this is incorrect. Because the measurement problems of identifying the effects of relatively weak treatments may be insurmountable, one should not write off the program as useless without further thought. However, there is this general presumption. So if you are a program manager who feels that his program is basically a good one but not very influential in the overall scheme of things, then the prospect of a good comparison group design not finding effects can be rather daunting.

Expense

It is sometimes argued that survey research in a comparative or experimental mode is too expensive and too time-consuming, relative to its advantages over "simpler" research procedures. This is somewhat difficult to address, because the detailed costs of most research efforts, experimental or not, are often poorly documented. I know of only one case where a substantive effort to compare the costs of alternative modes of social research has been made: that is, the accounting projects of the National Institutes of Education in the United States. As reported by Robert Boruch, randomization appears to have required much less than a one percent increase in research budgets, the increase being spent on payments to control group members and to experimental group members in return for their cooperation.

Of course, experimental or comparative survey research can be expensive in absolute terms. For example, the Manitoba Basic Annual Income Experiment which I mentioned cost approximately 25 million dollars. To take another example, the research budget for the Housing Allowance Experiments in the United States (excluding transfer payments to participants) was in excess of 130 million dollars and this figure does not include the substantial involvement of government staff and facilities in the research which one would have to include in a full costing. Good research on a question of national importance is not cheap. Whether it is "expensive" depends upon the potential savings from having a good program rather than a mediocre or bad program. There was a move in the United States Congress, at one point, to require that one and

one half percent of all monies appropriated for new programs be allocated for evaluation research activities. Roughly, this is about 100 times what the Canadian government typically spends on similar activities. Whether or not well designed survey research is likely to be cost effective depends upon one's judgement of whether the effectiveness of the program can be improved by more than that one percent of budget (or one 1/100th of a percent of budget).

Of course, survey research on some programs will be more expensive and time consuming than others. Those programs which are expected to have long-term effects or to have effects only after a long period of treatment can be particularly expensive to evaluate. Nevertheless, given a decision to evaluate this type of program, it seems reasonable to suppose that good design is better than a bad one. In the past, there may have been some association between the size of a research project and the rigor of its design. Larger and more expensive research may have been associated with more rigorous designs. But the direction of the causal relationship was probably from large budgets to rigorous designs rather than from rigorous designs to large budgets.

Ethics

In some cases, it will be inequitable, unethical, illegal, or otherwise imprudent to assign some members of a target population to a "control" (no treatment) condition. If the program under consideration is a demonstration or pilot program, then the comparison group is less likely to feel ill-used, especially if they receive some compensation for their participation in the experiment. On the other hand, if the program has been legislated already and is in operation, then eligible persons are likely to feel that it is their right to receive the program "benefit" and will likely feel deprived if assigned to a no-treatment comparison group.

One approach to this problem is to compare the relative effectiveness of different types of the same program, rather than comparing a single treatment with no treatment at all. For example, it is certainly possible to devise a number of incentive programs to encourage private industrial firms to undertake higher levels of research and development. Let us imagine that we devise three

such potential programs, and that the state of our knowledge does not allow us to judge with confidence which of the three programs would be the most effective. In such a case, applicants for assistance can be screened for basic eligibility and then randomly assigned to one of the three alternative programs. This will allow us to compare the relative effectiveness of the three programs, although not, of course, to measure the absolute size of the effect compared to no program.

There are other more sophisticated methods of overcoming the ethical problems which may be inherent in some randomizations. For example, if the treatment is a life and death matter, it is possible to assign people to the intuitively most attractive treatment until a failure occurs, whereupon one assigns people to the major alternative treatment until again a failure occurs, whereupon one assigns further people to the first treatment until a further failure occurs, and so on. There are a number of different randomization procedures which are appropriate to different program situations, and, in general, it seems possible to satisfy equity and ethics within a good research design. In fact, sometimes equity considerations can assist in the construction of a comparison group. A recent example comes to mind from the Department of Industry, Trade and Commerce. The Department has a program called PEMD (Program for Export Market Development). One section of this program assists Canadian companies to bid on projects overseas. On grounds of equity the Department decided that if there were several Canadian companies wishing to bid on the same project, then it would extend support to none of them, rather than to make the difficult decision to choose one over another. So when the Department subsequently examined the program to see whether the incentives had really made a difference to the behaviour of the companies involved - whether the grant really made a difference to them going ahead with a bid - then there was a reasonably good "no treatment" comparison group consisting of these firms rejected on equity grounds, whose actual behaviour in going ahead with the bid or not provided a basis of comparison with those firms who did receive a government grant.

Political Problems

The simplest explanation for the lack of control group designs in survey research in Canada may be that the users of the research do not want rigorous designs. It seems at least worth exploring this idea in conclusion to this paper. In the United States, the separation of powers between the adminstrative branch (the bureaucracy), the legislature, and the judiciary, provide a number of legitimate competing interests in the program. Specifically, a member or committee of Congress has a right to know about the programs within its jurisdiction - an "oversight" function - but has no direct responsibility for or authority over the bureaucracy which probably generated, and certainly adminsters the program. So there are institutionalized actors within the system with both the power to initiate survey research, and an interest in having it done rigorously, let the chips fall where they may.

In contrast, in Canada a minister has both a legislative and an administrative hat and so if he initiates data collection that may or may not be complimentary to his department, he is to some extent fouling his own backyard. This is also true of deputy ministers, perhaps even to a stronger extent because their career and reputation is even more closely tied in with the success of particular policies and programs. In this situation, ambiguity has a value. It is useful to know something about the program, but not too much or too precisely. The potential risks outweigh the potential rewards, especially given the general suspicion that many government programs do not make a great deal of difference to the problems they are meant to address, and certainly seldom achieve the grandiose goals which often form part of the program mandate. Essentially, the incentive structure for both senior civil servants and politicians favours large programs rather than effective ones.

If it is true then that it is not technical problems, expense problems, or ethical problems, that have lead to such a dearth of good comparative survey research designs, then perhaps we are correct in ascribing the failure to a structural problem of inadequate or perverse incentives.

There are at least three recent developments in the Government of Canada

which might change this situation: a growing pressure from central and service agencies, especially the Office of the Comptroller-General, for better quality control of policy-related survey research; the "envelope approach" to expenditure budget management which will encourage more vigorous inter-program comparisons; and the move towards freedom of information which will give the research efforts greater visibility, and bring them more closely under public scrutiny. Nevertheless, despite all this, the incentives of the main clients for survey research remain the same - that is, to produce information about the program which is sufficiently detailed to demonstrate knowledge and control, while being sufficiently equivocal not to pose a threat. The only way past this roadblock is to get new clients for survey research who have more detached interests. In the Canadian context, this probably means having one or more of the central agencies conduct evaluative survey research directly. There seems to be no immediate prospect of the Controller-General's office, or the Office of the Auditor-General, or the Prime Minister's Office, doing this.

Boruck, R.F. and Reicken, H.W., Experimental Testing of Public Policy, US Social Science Research Council, 1974; and, Social Experimentation, Sage, 1978.

Abt,C., The Evaluation of Social Programs, Sage, 1976.

RESUME

Ce document traite des avantages de concevoir les enquêtes de manière à pouvoir comparer les résultats avec des données existantes. L'auteur démontre qu'il est difficile dans la pratique d'évaluer les données d'une enquête unique en l'absence de données comparables, c.-à-d. en l'absence de données de contrôle ou autres repères.