300
EPMMFSP

INTER-OFFICE CORRESPONDENCE

FROM: BEA

DATE: July 7, 1984

TO:			COMMENTS:
	RWL		
	BEA		
	PHC	fyi	
~~BEA~~	AR		
~~[illegible]~~		file	

Pls. choose your first alternative (No M.J.S. report only).

I would be reluctant to give her the MIS report without: a) agreement of CBOs & AAI
b) full explanation that the MIS data does not record child care utilization and should not be used by DOL in lieu of their own reports

I gave Alexander the questionnaire but did not promise anything else.
PHZ

SUBJECT: MFSP Item

Our special meeting with the Advisory Committee on MFSP on Tuesday, July 10, was called to discuss the Hollister report. He offered a number of recommendations for modifying the research and evaluation plan, and after discussing appropriate next steps with LDS, it was agreed that the Advisory Committee, or whatever number of members might be available, should be convened to discuss the report before any major changes are made in the research plan. (The next regularly scheduled meeting of the committee will be on Thursday, October 11.)

ATTACHED

All committee members have received copies of the Hollister report, and have been asked to come prepared to discuss the recommendations, especially those related to introducing random assignment at some of the sites, and efforts to change sample size in order to strengthen the reliability of research findings. The Committee also received copies of the last MIS report, and volumes I and II of AAI's Annual Report. These documents, however, will not be discussed, except in passing, at the meeting. (These documents were sent to your office last week.)

The agenda will be relatively brief, beginning with a short introductory statement in which I will trace where we are in the mid-course review of the research plan, and continuing with a brief introductory statement by Hollister. Phyllis Wallace will then lead the Committee through a point-by-point discussion of the Hollister recommendations. The agenda, however, is not set in concrete, and we would be happy to have your views on an alternative approach to the discussion. (A copy of the Hollister report was sent to your office on June 2. His final report was forwarded this morning.)

Request from the Women's Bureau

As you recall, shortly after the MFSP program was publicly announced, we received an inquiry from the U.S. Department of Labor's Women's Bureau about possible collaboration in support of the venture.

FORM 106

After some discussion, it was agreed that collaboration of the type suggested by WB would not be appropriate, but that we would seek ways to cooperate in jointly achieving the goals of the program.

Since Fall, 1982, we have cooperated with the WB in several ways, most notably sharing with them information on the CBOs selected to participate in the program, and assisting in the formulation of their plans to help with child care support. In early 1983, the WB awarded contracts to four of the six CBOs (OIC, WOW, AUL and CET) for child care assistance. The grants, valued at $100,000 each for the first year, helped the CBOs meet their child care requirements under the program. During the second DOL contract year, ending in September 1984, the grants were reduced to $80,000 per CBO and we are now advised that during the third year, the grants are likely to be further reduced to between $45,000 to $55,000 per CBO.

Dr. Lenore Alexander has requested access to our research and MIS reports as part of our "cooperative" relationship. We have briefed her, in general, on progress of the program, but thus far, we have not shared AAI reports with the Women's Bureau.

I am somewhat hesitant to release to the WB full reports on the program because of uncertainty about their possible distribution throughout the federal government. Moreover, much of the current research information is still in the early stages of development, and premature release of such information to federal policy makers might not be advisable. From the very beginning of the program, there was every intention to share the research findings with a broad range of interested persons, but it was always understood that the data would be distributed only after careful preparation by AAI staff and prior review (but not censoring) by the Foundation.

One possibility for responding to Dr. Alexander would be to give her the last MIS report, which summarizes the enrollment data, including characteristics of applicants and participants, together with services now being provided by the CBOs. We can meet with her to discuss the meaning and interpretation of the report in order to avoid possible misunderstanding of the data. The alternative- of course, would be to send the WB the AAI Annual Report which summarizes both enrollment data, and critical research and methodology issues.

May I have your comments on how best to respond to the Alexander request? Thanks.

B.E.A.

EPMG FSP 200

JUN 27 1984

For Adv. Committee Meeting, Tues- LDS

	Pitz JUN 27 1984		Pitz	Female copy

Review of Minority Single Female Parent Program
Research and Evaluation

Robinson G. Hollister Jr.
Professor of Economics
Swarthmore College

29 May 1984

Introductory Remarks

This is a report submitted as part of a review of research and evaluation of the Minority Female Single Parent Program (hereafter referred to as the MFSP) which the Rockefeller Foundation asked me to undertake.

In the report which follows, I try to respond directly to the questions posed to me in the Terms of Reference provided by the Foundation and each section is devoted to one of the four main questions posed.

I ask the reader to understand that this review was carried out in a very short period of time and the report itself was written under extreme time pressure. Therefore, I ask indulgence for poorly elucidated arguments, infelicities of style and a lack of compression or summary in various sections. More important, the statistical calculations provided in the text were done extremely rapidly and have not been carefully checked by myself or anyone else. Normally, I would have someone check over such work for me, but time did not permit, so errors may persist.

In carrying out this review, I have examined Foundation documents on the project, reports by Abt Associates Inc. (hereafter, AAI) and correspondence between the Foundation and AAI.

I made brief visits to two sites, Wider Opportunities for Women in Washington and Opportunties Industrialization Center of Rhode Island in Providence, and discussed the program with the project directors and other personnel at each site. Each of these visits was about three hours in length.

I met with researchers at AAI in Cambridge over a two day period. The FSMP project director was unable to attend the meeting, but all the rest of the relevant staff participated.

I discussed the program briefly with the President of the Foundation, Richard Lyman and the Vice-president, Lawrence Stifel. I had longer discussions with Bernard Anderson, Director of Social Sciences and Phoebe Cottingham, Associate Director.

I also discussed the project with Professor Michael Borus, consultant to the Foundation.

I am acutely aware that a report based upon a quick review of materials and conversations with a limited number of individuals may contain many errors or misinterpretations of fact as well as oversights of critical issues.

Before turning to the questions posed in the Terms of Reference I would like to make two general points about the MFSP. First, one of the major purposes of this project has already been achieved, namely, through an action-demonstration program attention has been called to the problems of minority single parents and their families. There can be little doubt that the Rockefeller Foundation's efforts have played a

major role in heightening public awareness of these problems. Second, even if the project fails to provide conclusive evidence of significant impacts of these programs on the economic conditions of the participants, the data collected will provide a rich resource for continuing research on the problems of minority female single parents.

I now proceed to the questions posed by the Terms of Reference.

Responses to the Terms of Reference Questions

In trying to answer the questions posed by the Terms of Reference it is important to try to specify the audiences which the foundation seeks to address with the information drawn from the demonstration because different audiences have different standards of evidence. For the broad public the standards of evidence will be relatively low. A good journalistic account will suffice. Among those concerned with policy making there is a sharp division between those who will find descriptive material plus some rough quantitative measures of program processes--e.g. enrollments, characteristics of participants, lengths of stay in the program-- sufficient, and the group which would require more formal quantitative evidence of the effects of the program. Even within this latter group there is a gradation in what are considered acceptable standards of evidence, with a small group of technically oriented persons who carefully scrutinize evidence on program effects to be assured that what has been measured is an effect of the program and not simply an artifact of the process which sorted participants into the program. While those at this last extreme are relatively small in number, their influence can be great. It is my opinion that the standards of evidence have been rising over the past decade and that the proportion of those who would accept largely descriptive material as a firm basis for policy formation is diminishing, particularly in policy making circles. For convenience of exposition I will refer to three audiences which, I believe, the Foundation might wish to address. The first group I will call the "descriptive" group, the second group I will call the "technical" group, and the third group I will call the "highly technical" group.

Now let me turn to the questions posed in the Terms of Reference.

1. Given the current level of enrollment and the likely prospects for the remainder of the five year program what questions can most reasonably be answered concerning the impact of MFSP program and its participants?

In attempting to answer this question I will in general assume that the research will go forward according to the current research design including, in particular, the survey design option as indicated in the document, New Survey Design Option, dated the 27th of March 1984. In order to try to provdie a reasonablly full response to this question, I start with a list of questions that were posed in the research design, as reflected in the research design document written by Abt Associates, Inc. (AAI) dated 18 February 1983.

a. How do project applicants compare with project participants and with the national population of single minority group economically disadvantaged female parents?

b. What kinds of individuals are accepted for project participation? How do sites differ in their project selection criteria?

c. What happens during project participation, and how do sites differ in project implementation?

d. What are the effects of project participation and post-project services on the participants?

e. What, in general, are the effects of project participation and post-project services in comparison with persons not receiving such services?

f. Do the different projects have differing effects on participants and, if so, why?

I will now go through these questions and provide a rough judgment as to the degree to which they may be answered with the information which will be generated by the current research design, given the level of projected program enrollments.

a. How do project applicants compare with project participants and with the national population of single minority group economically disadvantaged female parents?

The answer to this question will depend entirely on the information coming from the MIS forms since it is only in these forms that data are gathered on project applicants. As currently designed, the MIS application form is filled out by persons who physically appear in the office of the Community Based Organization (hereafter referred to as the CBO). AAI tells me that, while initially there may have been some slippage in actually obtaining an enrollment form for every person who appeared to apply, they feel that currently they are capturing at least 90 percent of those who physically appear to apply for the program. The application form is fairly detailed so there are a reasonable number of items of interest for which comparisons between applicants and project participants will be readily provided. How easy it will be to match these data items to data on the national population of a single minority group female parents, requires a detailed examination of the sources for estimates for the national population and the precise character of the data collected in those sources (an effort which I believe is currently underway at AAI). I have had not had time to make such a detailed comparison of data items but I would guess that it will be possible to make gross comparisons with the national population.

b. What kinds of individuals are accepted for project participation? How do sites differ in their project selection criteria?

There will be two sources of information for purposes of answering this

question: the data from the MIS forms for those actually enrolled (that is both the items on the application form and those on the enrollment form) and the baseline interview questionnaire which is given to persons enrolled in the program. (It now appears, I understand, that because of the lower than expected enrollment rates, the base line interview is likely to be given to all persons enrolled in the program rather than some portion of those persons as originally anticipated in the research design.) These baseline surveys will make it possible to make detailed comparisons of participants across sites.

The question of how sites differ in their project selection criteria might be answered in a number of different ways, depending on the precise meaning of this question. For an example, a comparison of the MIS application forms with the MIS enrollment would give some indication of selection, but it would be a combined selection on the part of the program and self-selection on the part of the applicants who decide to proceed to participate in the project. Other sources of information must be tapped if one would seek to answer the question precisely with regard to selection exercised by the CBOs themselves. It appears that one has to depend upon statements by the CBOs themselves or the information gathered through the onsite researchers (hereafter OSRs). The structured logs that the OSRs have been given directs them, in the list of topics to be addressed specifically, to investigate the selection processes by which projects choose enrollees from among those who apply to the project.

This question also might be interpreted to apply to not just selection among the applicants of those to be enrolled, but also selection within the program for the different activities to which those enrolled are assigned. For this latter question primary reliance must be made on the reports by the OSRs. Of course, there are bound to be discontinuities between the stated selection criteria of the CBOs and what one actually observes from the data on participants. The combination of the OSR reports and the MIS material might allow one to determine precisely the degree of such a discrepancy, but to do so might require a substantial investment of analytic resources. It is not clear to me that obtaining precise answers would be worth the resources required.

c. What happens during project participation, and how do sites differ in project implementation?

The primary objective of the OSR segment of the research design is to provide a rich and somewhat systematic description of these processes. I would guess that this will constitute a substantial contribution to the research. I must state, however, that I am not personally experienced in using information collected by means of the procedures for the OSRs - although I am familiar with the fruits of participant-observer studies. I probed AAI somewhat on the question of how this information is to be used and they provided me with some examples from their past research using similar processes. I have not, however, had time to examine these materials in depth. I have no doubts at all that the OSRs will make an important contribution to the research in this project. However, it is not clear to me whether the

information they provide can be drawn together in any systematic format. AAI has tried to structure the reporting of the OSRs in such a way as to increase the likelihood of obtaining comparable information across the sites and with the different OSRs in any given site over time.

The description by the AAI analysts of the ways they have, in the past, used the type of information provided by OSRs to reformulate variables and models for their more formal analysis convinced me that this type of information could be useful for the more formal analysts.

Overall, then, I conclude that there should be an adequate basis for a rich description of what happens during project participation and how the problems of implementation differed across the sites. Whether this wealth of detail can be reduced sufficiently to yield convincing generalizations and comparisons is less obvious to me.

d. What are the effects of project participation and post-project services on the participants?

There are two general dimensions of this question: first, are the pieces of data which are necessary in order to measure any potential effects of the program going to be collected and, second, is the sample size likely to be adequate to measure effects of the program at a reasonable magnitude, given conventional standards of statistical significance?

With regard to the question of data to be collected, the baseline questionnaire, as now constituted, seems to provide adequate coverage of the most important areas (with one or two exceptions which I will discuss below).

Since we do not have a draft of the follow-up questionnaire, one cannot be sure how adequate it will be. If the questions asked on the baseline were simply repeated in the follow-up one would have a reasonable indication of changes in the demands for child care, employment, sources of income, and family structure. However, the follow-up questionnaire should be designed somewhat differently, I believe, in order to capture more systematically the events which occurred between the time of enrollment and the follow-up interview.

There is one problem which arises with respect to sole reliance on the baseline and follow-up survey to measure program effects, and that is that one cannot differentiate the changes that are due to the effects of the program on the participants from those changes that are due to other factors affecting their lives. This is, of course, why one would like to have a control(or comparison) group, a topic to which I return with the next question. In addition, this concern suggests that it may be helpful for the research group to try to gather systematic information from other sources as to changes in the environment in which these women are situated which might affect their circumstances and alter the impact of the program, for example: changes in the general employment situation in the area, changes in the availability of various government sponsored programs such as job training programs,

child care support, welfare requirements, food stamp availability, and transportation services. While the comparison group will provide some insights on the extent of such changes in external factors, it may not do so adequately. Thus some attention should be given to the extent of systematic data collection relating to these factors which needs to be undertaken to supplement the information provided through the comparison group. It is my impression that, to date, this has not been given serious consideration in the research.

With regard to the second dimension of the question which is sample size, it appears that this question has not been systematically addressed as part of the research design effort. There is no indication that anyone has asked: suppose that the analysis were limited to pre-program post-program comparisons for the participants alone how adequate is the sample size?

e. What, in general, are the effects of project participation and post-project services in comparison with persons not receiving such services?

This is the question to which the central impact analysis is addressed. As the same baseline and follow-up questionnaires will be utilized for the comparison group as well as the participants, the comments regarding the previous question apply equivalently to this question. Turning to the issue of adequacy of the sample to detect the program effects and their comparison with persons not receiving the services, there are a whole host of technical issues. I will only sketch these issues very lightly here.

First,there is the issue of whether the raw sample size is likely to be large enough to detect the impact of the program. In particular, has the lower than anticipated enrollment rate seriously threatened the adequacy of the sample? At the most summary level I would say that if the enrollment and survey numbers as reflected in the March 27th sample design are achieved the sample should be adequate to detect effects at a reasonable level for both the demonstration overall (i.e., pooling all the sites) and for four of the five individual sites.

This summary conclusion, however, should be underpinned with a bit more discussion. In order to make the kind of conclusion I have just stated, one must start with some a priori estimates of what the likely magnitude of the impacts of the program might be, or at least what is the lowest level of impact which would be relevant to the formulation of policy on the basis from information provided by the demonstration. Curiously, in none of the documents that I have seen thus far has there been a discussion of either of these magnitudes.

There are at least two ways of developing such a priori estimates. First, one can look at studies of previous programs dealing with this population (or as closely approximate to this population as possible) and the estimates of the magnitude of effects of such programs. Second, one could start from rough estimates of the cost of the program treatment and assume a rate of return on those costs which would be necessary for the program to be socially acceptable. (In both these

cases there are some complications regarding the rate of decay of the effects of the program which I won't go into here.) I must stress that I have not investigated either of these approaches in detail with respect to this project (although we did do so for the design of the Supportive Work experiment).

With respect to the first approach, looking for other program studies that appear most closely related to FSMP, the estimates of CETA training programs for women and the estimates of the Supportive Work program on the AFDC target population seem most relevant. Of course differences in program and population served between these two programs and the MFSP may be substantial, but I ignore that here. For both the CETA programs and for Supportive Work AFDC the order of magnitude of gain in earnings for minority females was in the range of 800 to 1100 estimated annual earnings. For Supported Work we can also obtain the estimates of the effects of the program on employment rates. The estimates indicated that Supported Work increased the employment rate of AFDC women by .07 above the employment rate for the control group AFDC women of .35.

Using the second approach to deriving an impact estimate, one could say if the costs of the program per participant were about $2500 and the required rate of return was 5%, then the expected minimum annual gain would have to be $125. If the cost per participant was $5000, the minimum acceptable return at 5% would have to be $250. I would guess that this approach to setting standards for expected impact would not be attractive to the Foundation since it specified explicitly in the initial design that there was to be no benefit-cost analysis and this approach is based upon a benefit-cost conception of research design. Therefore, I will not refer to this standard in subsequent discussion

In what follows, I use as crude benchmark figures for the evaluation of the sample design $800 for the expected earnings impact and .07 for the expected impact on employment rates.

Using these benchmarks we can look at the adequacy of the sample for detecting effects of the magnitudes suggested. There is an important question here of what is the appropriate way to look at the sample statistics. In the research documents and memoranda that I have read so far, all of the presentation of the sample size options and their likely adequacy have been stated in terms of the minimum detectable program effect. This way of evaluating sample adequacy has come to be regarded as inappropriate for the design of experimental or evaluation samples. In statistical terms it is essentially a discussion of the probability of what is called type 1 error. Concern with type 1 error is appropriate for the analysis <u>after</u> the data have been collected, but for the purposes of <u>a priori</u> sample design the appropriate criterion is is the degree of type 2 error, and the appropriate analysis for that is something that is referred to as the statistical power of the sample. In an appendix I develop this in a little more detail (when I spoke to AAI last week, they were already in the process of developing power estimates for the different sample options, apparently at the direction of the new project director Steve Kennedy). One way to state what the power estimates tell you is as follows: if the true impact of a

program like this were of a given magnitude, say $800 in annual earnings, what is the probability that the sample of observations you collect would be such that your statistical tests will lead you to conclude there was an impact statistically significantly different from zero?

I have done some very quick calculations by hand of the power provided by the latest sample design, using as a starting point previous estimates by AAI of the minimum detectable response. They are reported in the table which follows.

ESTIMATES OF STATISTICAL POWER
FOR THE SAMPLE DESIGN OF MARCH 27, 1984

SITE	MINIMUM DETECTABLE RESPONSE	TRUE EFFECT 90% POWER	TRUE EFFECT 80% POWER
	EMPLOYMENT RATE ASSUMING AVERAGE RATE = .3		
ATLANTA	.061	.101	.087
BROOKLYN	.076	.125	.108
PROVIDENCE	.065	.108	.094
SAN JOSE	.061	.102	.088
WASHINGTON	.069	.115	.099
POOLED	.029	.048	.042
	ANNUAL EARNINGS (IN DOLLARS) ASSUMING STANDARD DEVIATION = $2453		
ATLANTA	324	537	473
BROOKLYN	415	687	606
PROVIDENCE	348	576	508
SAN JOSE	338	560	493
WASHINGTON	368	609	537
POOLED	156	258	228

(Before discussing this table I want to note that I have used the estimate for the standard deviation of earnings which AAI used in its recent documents. However, AAI's notes indicate that this was calculated from the variance in total money income for female headed families as calculated from the Current Population Survey. But total money income as defined in the CPS includes welfare payments as well as earnings, therefore it is not a precise estimate of earnings variance. Whether the appropriate variance of earnings would be larger or smaller is not clear to me a priori, but steps should be taken to obtain better estimates of the appropriate magnitude. I have not had time to do this myself, so I have used the AAI estimate in my calculations for this report. I have chosen to use .3 as the average expected employment rate in estimating the standard deviation for the employment rate, since that rate was reasonably close to that found for the controls in the Supported Work study.)

If all the sites can be pooled together for analysis the current sample provides very good statistical power for either the employment rate or earnings. Even if the overall effect were as small as .048 difference in employment rates or $258 in earnings, there is a 90% chance that with this sample the results would yield a finding of a statistically significant program effect.

When we look at each site individually, to see if it will be possible to obtain estimates of impact for each site taken separately, we get different stories according both to site and to which outcome measure (employment rate or earnings) we are concerned with. For earnings, all the sites seem to have a high probability of being able to detect an earnings differential if the true impact is as large as $800. For the employment rate outcome, the margin of power looks more questionable. If the true impact were .07 difference in employment most sites would have only about 50% power, that is a 50% chance of reaching a conclusion that the program impact on employment rates of the program were significantly different from zero. (The value labeled minimum detectable response is equivalent to the value for which the sample has 50% power).

Since the power estimates for the earnings impact seem quite adequate, however, I will choose to ignore the weakness with respect to the employment rate outcome. However, I have made all these calculations very rapidly and have questions about the underlying earnings variance estimate, it is important for AAI and the Foundation to pursue this work in more detail immediately.

While in general the samples appear adequate at the site level, I do have questions about the Brooklyn site. These arise because this site has the lowest sample, the lowest power and because from what I have seen in the documents, may be the weakest program. In addition, some of the researchers have expressed skepticism about the ability of the program to reach even the low numbers called for in the sample design. I will return to a discussion of Brooklyn at a later stage in this report.

Overall then, I conclude that, looking just at the raw numbers in the

sample design and expected outcomes, there is a reasonable chance of detecting differences between participants and the comparison group. There are several considerations which strongly condition this conclusion, however. They are: i)considerable turbulence in the first year at several sites, ii) the problem of selection in the participant group, iii) the adequacy of New Haven as a source of comparison observations for Providence, iv) the desire, or necessity, of looking at sub-groups within the participant group. I will discuss each of these considerations in more detail, but as a general principle, one can think of each of these considerations as eliminating some portion of the sample and thereby reducing the effective sample size and statistical power for the outcome measure of interest.

i) I have been told that several of the sites had considerable difficulties in getting the program underway in the first year and this is documented to a degree in the AAI reports on implementation. Such first year turbulence is to be expected in any program, even if implemented by an on-going organization such as a CBO. One could take the view that from the point of the research this is irrelevant: the impact study is to measure the effects of the "black box" MFSP without regard to specific implementation problems. However, most observers would argue a more appropriate procedure is to attempt to measure the impact of the program during a period in which it is reasonably stable, or at least to analyze results separately for a period after the initial start-up.

Suppose, for example, it were decided that difficulties in the Providence site were such that the analysis should focus only on observations for the second year onward. The effect of this on the power of the Providence sample would be as follows: whereas if the full three year sample were used the sample would have 90% power with a true impact of .108 difference in employment rates or $576 difference in annual earnings, with just the sample after year two the power would fall to 64% (if you prefer minimum detectable response it would rise from .066 to .091 for employment rates and from $348 to $486 for annual earnings).

There are two different ways of thinking about this problem: if you include the first year you may expect the average impact to be smaller because you have included a period when the program was relatively ineffective on the average, or alternatively, if one excludes the first year observations, the effective power of the sample has been reduced. Either way, one's doubts about the adequacy of the sample for detecting the impact are increased. To proceed further with this consideration one would have to do more detailed and precise analysis of how much first year turbulence might reduce effective sample in each site and I have had neither the detailed information nor the time to go further with this.

ii) The problem of selection in the participant group has been recognized from the outset as the most serious threat to the validity of any estimates of the program on participants in comparison to persons not receiving such services; once random assignment to the program and a sample of controls was ruled out, any method of

developing participant-comparison contrasts would have been subject to this threat. AAI, the Foundation and its Advisory Council have all recognized this problem from the outset but have decided to proceed with the current strategy. I can contribute only a few observations, my personal assessment about the seriousness of the threat and some suggestions for further steps which might help to reduce somewhat the chance that the final analysis will be seriously vulnerable to criticism on these grounds.

My concern about the selection bias focuses primarily on motivational self-selection. In both of my brief visit to sites (Washington and Providence) program operators stressed the very serious problems of getting members of the population into skill training. The MIS data show a very substantial difference between applications and enrollments (from September to December of 1983, 41% of applicants were enrolled within the same period). AAI has been seeking to establish a "uniform enrollment" point at which the baseline survey questionnaires are administered and this apparently occurs after the applicants have engaged in some substantial activity at the site (in at least one case, I believe, they will have to have shown up for three separate application, interview-testing, and enrollment appointments in order to be considered enrolled and given the baseline interview).

This is all quite reasonable from the point of view of the program and from the point of view of the survey interviewers, but it seriously deepens, I believe, the potential problem of selection bias. If, in this population, strong motiviation is required to get people to apply and even more so to show up to enroll, and this type of motivation is also important in getting and holding a job, and if this motivation can not be strongly correlated with those characteristics measured in the survey, then the problem of selection bias in the contrasts of participants and comparison group members is quite serious. In addition to self-selection on the part of participants, there is undoubtedly some degree of selection among applicants on the part of the CBOs themselves, and the degree of this type of selection is likely to vary from site to site.

Some may argue that there is not much that can be done and, indeed, not much need be done about the selection problem. There have been a lot of studies of programs, and especially training and employment programs, which have used comparison groups constructed in a similar fashion. There is a substantial part of the public and even most of the audiences I have label as "descriptive" and "technical" who would consider these problems of selectivity bias as irrelevant or too arcane to be of interest.

This argument has some merit. The audience which would be persuaded by criticisms about the validity of results from FSMP based on selection bias in participant-comparison group contrasts is the one I have labeled "highly technical". It is not clear that any further efforts to deal with this selection problem, special analyses or models, will be sure to convince them that the selection bias has been removed.

On the other hand, it must be considered that the Foundation has made a substantial investment in the part of the research devoted to creation of a comparison group and the surveys administered to them and the participants in order to measure impacts by contrasts between participants and comparison group members. This investment is put at risk by vulnerability to criticisms of substantial selection bias. Therefore, I urge that the Foundation and AAI take this problem more seriously than they appear to have to date and investigate further steps they might take to reduce the likelihood of selection bias or to correct for it in the analysis stage.

While AAI discussed the selection bias problem in their research design document and sketched out the techniques they might use to reduce it, they need now to try to specify much more precisely exactly how they are going to try to model the selection process. There is by now a fair amount of experience with trying to use the selection bias modeling techniques they propose in their design and this experience is not entirely reassuring: sometimes the selection modeling seems to work well, sometimes it is clear it has not worked at all and many times it is between these extremes. As far as I know, at this stage, analysts have not been able to specify the types of conditions which are likely to generate success or failure with these methods. I can make a few comments,however, based on my limited experience with the use of these sorts of techniques and with the literature on them.

As I considered the problem of the substantial self-selection bias problem based upon motivation, I tried to think of the most closely comparable situation. The Jobs Corps evaluation seemed to come the closest. Because of the residential nature of the program, it is likely that a special motivation was necessary for a youth to become enrolled in the Jobs Corps. The evaluation by Mathematica Policy Research used a comparison group method somewhat similar to that of FSMP. That evaluation has been heavily scrutinized and often reviewed by other experts and seems to have survived critiques pretty well. The techniques used were similar to those suggested by AAI. A key feature, however, was the inclusion of two variables which helped to identify (in the mathematical sense) the equation used to estimate the probability of selection into the Jobs Corps. Those variables were the geographical distance from a Jobs Corps Center and the percentage of youth in the "neighborhood" who had been recruited in recent years.

I draw two points from this consideration of the Jobs Corps example. First, Jobs Corps, perhaps, provides a counter-example for those who argue that the "highly technical" audience is unlikely to be persuaded by any "fix up" of selection bias problems. Second, the success of the "fix up" may have depended on the availability of two specific variables and it is not clear to me, that AAI analysts trying to model selection for the FSMP processes will have any clearly identifying variables that can play the strong role such variables did in the Job Corps analysis. This observation suggests that it is important that the AAI analysts get to work immediately to

specifiy in detail how they hope to estimate their selection models. This is particularly important because if there are candidates for identifying variables, the data needs to be collected on such variables starting immediately. (In this regard, the analysts might take advantage of the work of the OSRs, whom, I note, have been specifically instructed in their structered logs, to try to observe and record the process of selection of enrollees from among applicants).

Another step which should be considered is the inclusion in the sample survey of some of the applicants who do not enroll in the program. These non-enrollees should be given both baseline and follow-up questionnaires. While it is true that the MIS applicant questionnaire gives a fair amount of data on non-enrolling applicants, having the fuller range of data provided in the baseline plus the follow-up may enhance the analysts ability to measure the degree of selection and attempt to correct for it. This need not be a full sample of non-enrollees, but just enough to provide some basis for analysis of the selection process.

I believe that consideration should also be given to the inclusion in the baseline of short forms of some of the cognitive and skills tests that are used by the programs for the assessment procedures. There are several arguments for the inclusion of such tests. First, if the CBOs are actually using such tests either to select enrollees from among applicants or to assign them to specific activities (and my impression is that some sites rely quite heavily on such tests), then to have some measures of this sort for the comparison group will considerably strengthen the ability to remove selection biases. Second, such tests may also help to capture some of the differences in what I have thus far called motivation, e.g., those with lesser cognitive or skill development may be more reluctant to undertake courses and training. Third, changes between baseline and follow-up in such test scores might themselves be regarded as outcome measures for the impact analysis. It appears that one of the major findings of the demonstration to date is the seriousness of the need for substantial educational remediation in the programs' target population as a precondition for improved employment and that the programs have been devoting substantial portions of their efforts to this end. In light of this, the impact study design should try to provide the means to measure the success of such efforts rather than depending solely on the much later appearing employment effects. I must hasten to add that I have relatively little experience with the administration of such tests as part of a survey instrument so I do not know how feasible it would be. The fact that the follow-up interview is to be by telephone may preclude their inclusion. My suggestion is simply that these possibilities be fully investigated. (Perhaps they were already considered at the questionnaire design stage and I have simply not been made aware of it).

iii) The adequacy of the current sample for drawing impact conclusions by participant-comparison group contrasts may also be limited for Providence because of the use of New Haven as a source for the comparison group. Differences in local labor markets, the

availability of child care, the features of the transportation network, the operations of the welfare program, the nature and availability of other government programs, the peculiarites of Providence as the capitol of a "city state" are all factors which could limit the comparability of participant and comparison group experiences. I realize that stating concerns about this is little more than second-guessing a decision made by the Foundation and AAI after serious consideration of alternatives. An immediately critical step would seem to be an analysis of baseline interviews already taken in Providence and New Haven to test for the degree of similarity of background and experiences at the outset. In addition, some attempt might be made to gather other information on the factors cited above for the two cities. I believe that serious consideration should be given to doing some comparison group sampling from the Providence area. While there may be a 50% chance that those sampled for comparison may eventually become applicants or even enrollees, I do not believe that this represents an unacceptable risk to take in order to get a better fix on the degree of non-comparability of New Haven.

iv) The adequacy of the current sample design must also be assessed in terms of the ability it provides to analyze sub-groups within sites. The analysis of sub-groups in relation to their equivalents in the comparison group will be desired, first, in order to try to determine whether the program is more effective for women with particular sets of pre-program characteristics, e.g., extent of education or previous work experience. The calculation of the effects of this on the power of the sample for a given outcome is straight forward. (AAI provided the Foundation calculations of this sort for the minimum detectable response). For example, for Washington a 50% subgroup of both participants and comparison group would have only 63% power instead of 90% power if the true impact on employment rates were .115 or the true effect on earnings were $609 (equivalently, the minimum detectable response would rise from .069 to .098 for employment rates or from $368 to $520 for earnings).

Many would argue that sub-groups must be estimated in order to determine whether participants in different types of training within the site had different levels of impact in contrast to their comparison group equivalents. This seems particular important with respect to a program such as that in Washington in which there is considerable separation in the types of training of the participants. This type of analysis, however, presents potentially severe selection bias problems similar to those outlined above. There is a potential for both motivational self-selection and program selection on variables not measured in the comparison group survey. The basic problem is that one has to have the means to sort within the comparison group those who, under like circumstances, would have selected, or been selected into, the given program components. The OSR research should provide some useful insights on these selection processes. Most of the suggestions outlined in the previous section would also apply to this problem. Some attention should be given immediately by the analysts at AAI as to whether and precisely how they intend to carry out such sub-group analyses. Again, to the

degree sites are using test assessments to sort participants, it is important to consider the feasibility and desirability of including tests in the baseline questionnaire.

To summarize the discussion of this section, I conclude that the sample sizes should be adequate to detect differences between participants and the comparison group at reasonable levels for most of the individual sites if one is concerned simply with gross differences in, say, employment rates or earnings including all three years of observations. Important caveats must be appended to this generalization, however. Ability to draw convincing conclusions regarding the effect of the program on participants in comparison with those not receiving such services may be seriously undermined by: turbulence in some sites in the first year program, serious problems of selection bias in contrasts of participants and comparison group members, the adequacy of New Haven as a source of comparison group members for the Providence site and the necessity to analyze subgroups within the sites. I have suggested a number of steps which might be taken to reduce the vulnerability of the research analysis to these threats.

f. Do the different projects have differing effects on participants and, if so, why?

The adequacy of the current research design to answer this question has been touched upon at several points in the discussion above. To discuss the question more precisely it is useful to break it into two parts: i) statistical tests of differences across sites; ii)other methods of estimating differential program effects.

I have already noted that for the most part the sample should be adequate to estimate effects of the program on participants as contrasted with comparison group members separately for each site and to detect statistically significant impacts of a reasonable magnitude. This means that there should, at least, be point estimates, e.g. the CET program in San Jose is estimated to have increased the annual earnings of participants when contrasted to the comparison group by $600. However, if one wishes to perform a statistical test to determine the estimated program impact in one site, say $600 in San Jose, is signficantly different from that of another site, say $800 in Washington, the power of the samples is greatly reduced. There are two reasons for this, first, the difference in impact between sites of, say, $200 dollars is much smaller than the difference within a site between participants and the comparison group of, say, $600 and therefore more difficult to detect. Second, and more important, the variance of this difference is approximately twice as large as that for the estimate of the single site effect. I would conclude, therefore, that the chances of establishing that the program impacts were statistically significantly different across the sites are quite small.

There are certain cases in which a statistical difference in site impact might emerge. Suppose for example, that the San Jose site was estimated to have an impact of $1000 and all the other sites were estimated to have no impact, a statistical test would sustain the

hypothesis that there was a significant difference in impact. However, if the individual site effects are small or the impacts are sizeable but of a similar magnitude across sites, it is very unlikely that site differences could be established as statistically signficant by the conventional standards.

ii) A second way in which one might seek to establish differences in impact by program type would be to try to classify programs according to more general characteristics and then test whether these characteristics were associated with significantly different impacts. AAI has already begun to try to classify programs according to various characteristics as part of its implementation research. In terms of establishing statistical significance of the impact of differences in program this approach has a probability of success which is essentially no different from that of the one discussed in the previous section, since the classifying device is basically a form of grouping sites.

One could try to push this approach somewhat further if one used MIS data to attribute at the individual level exposure to different types of program elements, e.g., the amount of exposure to basic education or the amount of exposure to counseling. This approach would increase the chance of finding statistically significant differences in impact according to program characteristic but the conclusions would be subject to serious reservations about selection bias equivalent to that described with respect to sub-group analyses in section e iv) above. However, all audiences except the "highly technical" would probably accept such findings as providing useful evidence about "what works".

There will, of course, be the rich data provided by the OSR work and analysts will certainly try to draw conclusions from these as to the program characteristics which "caused" any observed differences in site impacts. The "descriptive" audience will largely find this type of conclusion persuasive but the "technical" and "highly technical" are more likely to regard them as plausible and interesting hypotheses rather than conclusions.

Thus far I have discussed the questions posed in the research design document. There are three additional questions which I believe deserve some serious attention: child care as a barrier to employment, transportation as a barrier to employment and inadequate education as a barrier to employment. From my discussions at the sites and with the researchers, it seems that these are emerging as major issues and it is not clear that the current impact research design will provide analysts with adequate means to determine quantitatively how serious these problems are in the various sites. I would simply pose the question to the researchers (and Advisory Board members): suppose that in a given site any of these three - child care, transportation, education (or combinations of them) -were the principle reason that the program failed to have an impact, how would the data collected on the participants and controls permit analysts to determine that was the case? It is true that the baseline interview does list lack of child

care in several places for not getting a job or for leaving a job, but I wonder whether that will in fact be very useful. Many analysts are very skeptical of answers to general questions of that sort when given lists of reasons (health limitations is another such reason often given and viewed, skeptically). Perhaps the answers to these questions are straight forward but they are not self-evident to me. Except for the education issue (for which I have already suggested the inculsion of some cogntive and/or skill tests in the questionnaires) I have no immediate suggestions for how one might attack these issues. I can only urge that an attempt be made to develop some approaches.

There are another set of issues which I do not believe the research will be able to deal with effectively, for which I have no quick solutions to suggest, but which might deserve some further consideration by the Foundation and the researhers. A given program site might fail to have an impact for at least five general reasons: the program model was wrong, the particular CBO performed poorly, the business cycle was working such that there was little hope of employing these women, the cuts in Federal programs removed services essential for employment of this population, the nature of the population is such that these types of efforts cannot be effective. To what degree would the data collected in the research allow analysts to pin down the relative contribution of these five factors to a given site's failure to have a detectable impact? The research of the OSRs may give some insight regarding CBO performance but it will be quite diffficult to go beyond suggestive evidence of this type and convincingly assess the role of this and other factors.

2.I now return to the second question posed in the terms of reference: <u>how should the research design for the MFSP program be modified to assure the production of useful and policy relevant information based on the five-year demonstration project?</u>

I do not recommend really major changes in the research design of the project, at least as I would define the term "major changes". I do recommend consideration of at least two substantial changes and a host of smaller ones. The two substantial changes have to do with the Providence site and the Brooklyn site.

a. I believe that two factors which I discussed in sections 1 e. i) and 1. e. iii) pose a substantial threat to the usefulness of the impact evaluation in the Providence site. These are: substantial program turbulence in the first year and the use of New Haven as the sole source for the comparison group sample.

I have no doubt that those familiar with that program would argue that the observations from the first year would not represent a "fair test" of their capabilities or the "program model". If this view is accepted, it means substantial reduction in effective sample size for this site.

I have serious doubts about the validity of New Haven as a comparison site. These doubts might be relieved by an analysis of the baseline

data gathered thus far which showed no significant differences in the participant and New Haven comparison characteristics at baseline. Even if that were the case it would not be entirely persuasive since, perhaps, the most important differences are how employment opportunities evolve in each market over the next few years and the availability of relevant support services in the two cities.

As I suggested above, I believe it is important for the Foundation and AAI to consider the possibility of sampling for the comparison group from Providence and surrounding areas. It should also be considered whether one might sample at rates greater than one comparison to one participant. This procedure is sometimes referred to as an "unbalanced sample design" and I will return to discuss it in more detail below. I realize that both of these steps involve risks of original comparison group members becoming enrolled at a later date in the program, but I question whether these risks are overwhelming (for reasons I'd be happy to discuss in detail later).

In any case, an immediate look at the first year baseline participant-comparison contrast for Providence and New Haven is imperative.

b.I suggest serious consideration be given to stopping all survey interviews - both participant and comparison group - for the Brooklyn site and that the survey resources that would have gone into these interviews be reallocated to, at most, one or two other program sites.

I make this suggestion with considerable trepidation because I have not visited the Brooklyn site and because it is based more on intuition than on technical analysis.

The intuition is based on the following considerations. First, going back to the table on the statistical power of the sample, it can be seen that the power of the planned sample is lowest for the Brooklyn site: for 90% power the true impact of the Brooklyn program would have to be an increase in employment rates of .125 points or an increase in earnings of $687. Most of the reports to date suggest that this may be the weakest program,the one least likely to obtain such substantial gains. In addition, researchers have expressed some skepticism about the sites ability to meet the enrollment targets even as outlined in the 27 March 1984 sample design. Finally, the average length of stay of participants in the program was longer for Brooklyn than for any other site which may indicate not only difficulties in placement but also that the length of post-program follow up captured by the 24 month interview wil be much shorter.

What might be gained by reallocating the survey resources to other sites? To give just one example: suppose that the resources for 400 (50 participants plus 100 comparisons from year 2 and 250 combined participants and comparison from year 3) observations were shifted from the Brooklyn site to provide 400 comparison observations for the San Jose site. The minimum detectable response for employment rates for San Jose would fall from .062 to .052 and the true response necessary for 90% power would be .086 rather than .102. More importantly,

perhaps, the added sample in San Jose would provide much greater gains in power for analysis of subgroups. For a 30% subsample of both participants and comparison group the minimum detectable response would fall from .112 with the present design to .094 and the true response needed for 90% power would decline from .185 to .155. (What is suggested here is an "unbalanced design" for San Jose with 364 participant observations and 771 comparison group observations).

The basic rationale underlying this suggestion is to try to concentrate the resources devoted to the impact analysis (as opposed to OSR or MIS resources) in those sites in which there is the greatest expectation of positive impact outcomes. Since the primary purpose of these resources is to provide much stronger, more convincing evidence of outcomes than the other types of research can give, it makes sense to try to apply these resources most carefully. (These are the same grounds on which it is suggested that impact surveys not be made during the start up phase of projects)

My hesitation in suggesting this shift of resources arises from the problem of making sound judgements about which sites have the greatest likelihood of positive impact outcomes. There is concrete experience with this problem with respect to the Supported Work program. At about the middle of that demonstration we, the researchers, asked the staff of the Manpower Demonstration and Research Corporation, which was in intimate contact with all the sites, to rank the sites in terms of expected performance. After the follow up results were in we found no relationship between the estimated impacts of the various sites and the rankings given by MDRC. This illustrates, I believe, that even those with detailed knowledge and experience in providing technical assistance have a hard time "picking winners".

Note that removing the impact resources from the Brooklyn site need not imply that the program resources or the other research elements should be withdrawn. The OSR and MIS data would still provide some means to evaluate the effectiveness of this program.

I believe there are sufficient grounds at least seriously to discuss the possibility of the reallocation of survey resources.

c.I have been struck by the fact that none of the research design documents I have seen discuss the possibility of "unbalanced sample designs". I have suggested, in the two previous sections, two situations in which such designs might make sense and the Foundation may wish to consider the possibility of such designs in general. An "unbalanced design" contains substantially different numbers in the participant and comparison group. There are a number of reasons why one might wish to consider such designs and I will not try to review them fully here. A couple deserve mention, however. One thing about evaluation research which often frustrates program sponsors is the "long wait to get the follow up results". This long wait is usually dictated by the limited ability of programs to build up their numbers of participants rapidly and the necessity of obtaining sufficient sample size to obtain adequate statistical power. By drawing substantially more comparison group observations one can increase the

statistical power without waiting for program enrollments to accumulate to very large numbers. The problem with this strategy is that the statistical power added by putting, say, 10 more observations in the comparison group is not as great as that obtained by putting 5 in the participant group and 5 in the comparison group. Thus, in terms of total observations, it is more costly to obtain a given statistical power with an unbalanced design than with a balanced design. The trade-off, then, is between the greater survey cost of information gained in this way against the shorter time until follow up results are obtained. For example, the Foundation could consider unbalanced designs in the second program year as a trade-off with continued survey sampling in the third program year.

If one looks at the cost of the research information provided not just in terms of the research costs but the combined research and program operations costs, then unbalanced designs can be very cost-effective. To obtain a participant observation one must incurr not only the survey costs but also the program costs. Thus, in term of the total budget, it is much cheaper, up to a point, to obtain statistical power in a sample by adding low-cost comparison obervations than by adding high-cost participants. It is my experience that program sponsors never find this point of view attractive. But I feel I should call it to the Foundation's attention, just in case it has not been seriously discussed thus far.

d. I suggest consideration be given to dropping the postcard tracking system. After trying this system in Supported Work and analyzing its cost-effectiveness, we decided it was not worthwhile. It is true that the follow up interviews for Supported Work were only 9 months apart so the gain from trying to maintain interim contact may have been less. However, the early returns on the postcard experience with MFSP seem to me to suggest it is of questionable value.

AAI has been supplementing the post card tracking with telephone follow ups. If this effort is to be continued, I suggest that consideration be given to some partial data gathering as part of this telephone contact. The major cost of making a telephone contact is already being incurred, why not pick up interim data on a small number of significant items?

e. I have already suggested above a number of changes in the questionaire and I have a few more details to suggest which I will not include in this report but will pass on separately. One minor suggestion that could be important, however, is that the Social Security number of both participants and comparison group members be obtained in the baseline questionaire. At present these are collected for the participants on one of the MIS forms but they are not collected for the comparison group members. The Social Security administration will provide data over long periods of time on the level of earnings reported to Social Security as long as the request combines the observations in groups of 10. This type of data has been used increasingly in the past 10 years to provide both pre-program and follow up data for research on employment and training programs. While it is well known that there is not full reporting of earnings to Social

Security, the data have proved useful for many participant-comparison group contrast and could provide an inexpensive means of follow up beyond the 24 month interview. (I have already passed this suggestion on the AAI).

f. I suggest the Foundation reconsider the possibility of utilizing random assignment to program participation or to the comparison group in at least some of the sites. I realize that this suggestion will be irritating to many who consider this decision well behind them, but I feel I would not be properly fulfilling my responsibility as a consultant on evaluation research if I did not make this recommendation.

It should be clearly understood that the major gain from random assignment is not a gain in statistical power per se (there is some gain, but that is not the major contribution). The gain comes from the fact that the question of selection bias in the estimated impact of the overall program is removed. The Foundation must realize that the findings of the impact analysis may be subject to serious criticism because of the selection problem and to that degree its investment of resources in that aspect of the research remains at risk, even if attempts are made along the lines I have suggested above to improve the understanding of the selection process and to try to remove the bias.

There is, by now, considerable experience with the use of random assignment as part of the research in the evaluation of employment and training programs. While the program operators invariably resist this procedure, experience shows that they can live with it and that has not seriously hampered their program development.

There are a number of ways in which the problem of establishing random assignment can be approached which might reduce operators' resistance to the procedure. I will not try to detail those possibilities here, however.

The Foundation might at least explore whether some of the program operators are sufficiently confident of their operations and have a sufficiently large pool of potential applicants that they would find the use of random assignment as a rationing device acceptable.

g. While benefit-cost analysis was explicitly precluded from the original request for proposal to the research contractors, the Foundation has recently decided to explore the possibilities of developing some estimates of program costs. This development seems to me inevitable. Should estimated positive impacts from one or more of the programs emerge in the research findings the question from those interested in policy will certainly be: yes, but what was the cost of per participant of producing those impacts?

The cost analysis proposal which the Foundation received from AAI was the subject of a number of negative comments about AAI's work. I was abit surprised when I read it then to find what seemed to me not a grossly incompetent piece. From my conversation with the Foundation staff and with AAI, it now appears to me that the problem was one of

serious miscommunication between AAI and the Foundation about what was looked for. It is clear that AAI was offering a maximum cost analysis proposal while the Foundation was looking for a minimum cost analysis proposal. I believe that the Foundation's expectations of what it might take in resources to do even a minimally useful cost analysis was too low and AAI's expectations of what level of effort on cost analysis could be supported within the total context of research on this project was foolishly out of proportion.

The most important question is how best to proceed in this domain. It seems to me that no sensible proposal can be made without first determining the character and condition of the accounting systems of the sites; how much of an effort it will take to obtain even minimally useful cost data will depend critically on how much can be readily gleaned from existing site records. AAI researchers told me that they had met total resistance from the site operators when they tried to inquire (through their OSRs I believe) about the extent of cost data available. The Foundation will have to seek the cooperation of the sites in order to obtain even a realistic proposal about what it will take to accomplish a cost analysis. In some cases, existing site records may lend themselves quite readily to a quick cost analysis, in others, it may take a totally new data collection effort. In any case, it must be anticipated that where substantial services (e.g. day care, training, counseling) or subsidies are provided outside the immediate framework of the program, special efforts will have to be made to capture the full costs. For these reasons, I suggest that the Foundation fund a small exploratory study to determine the feasibility of developing cost estimates at least for some of the sites.

3. The third question posed in the terms of reference is: <u>What lessons from the MFSP program are likely to be most useful as a guide to future program planning and implementation as regards community-based efforts to help improve the economic status of single parents?</u>

Given that to date I have visited only two of the sites, and those only briefly, and that I have simply not had sufficient time to consider this question in depth, I am not going to try to answer this question at this point. I will be happy to attempt to do so at a later date.

4. The fourth and final question in the terms of reference is: <u>If significant revision of the research effort is recommended, how should the Foundation manage the research process so as to assure reliable results?</u>

I will set aside the question of whether the revisions I have recommended in the answer to terms of reference question 2 are substantial and simply comment on the general question of the future management of the research process by the Foundation. There are two parts to this question: first, what should be done about the research contractor, and, second, what about the management of research on this project within the Foundation itself?

While AAI does appear to have been negligent in certain aspects of the research effort, I think it would be too risky to try to change research contractors for any significant portion of the a research effort. While I could spell out my appraisal of AAI's performance in more detail, I don't think it would be fair (since I have not had the opportunity to discuss the project with Stephanie Wilson) or useful for me to try to do so at this point. However, let me give a few comments on each of the major elements of the research.

The work of the OSRs is the most certain, I believe, to provide useful material for the Foundation about how these programs deal with this population. That is not to say that this sort of evidence will establish the effects of the programs but rather that one can be fairly confident that what they produce will be useful to those interested in program processes. AAI seems to have a fair amount of experience in using this kind of research procedure and I doubt that there is another research contractor who would be likely to handle it better (though I must state again that I do not consider myself an expert on this style of research).

I think there are problems with the MIS design and have my own views on how it might have been done better. However, I believe it is too late in the development of the project to consider really major changes in the MIS. All parties have already paid a considerable price to get this particular system going and, though the sites complain about the heavy reporting requirements, the system does seem to be settling down so that it may be relied upon to provide useable information over the life of the project. I believe it is better to carry forward with this system than to try to turn back and develop a major alternative. My generalizations on the MIS are based on a very superficial review, however, and, while, I have some experience with MIS systems I again do not consider myself an expert. (This is an area in which I would say beware of the experts, however). I will provide more analysis of and comment on the MIS if the Foundation wishes. For the Foundation's information, I do want to point out that I regard the turn around time (time from the submission of forms to the production of summary tables) which AAI acheives for the MIS to be extraordinarily short. Processing these complex forms, making sure the data are reasonably clean and producing tables would, in my experience, normally take a good deal longer, so AAI is performing quite well in this regard. Undoubtedly, this fast turn around has been achieved at the cost of thoughtful analysis of the MIS data, but the Foundation should realize that that was bound to be a trade-off which had to be made.

For the impact analysis, there are some important shortcomings I've noted above, but I believe the Foundation is more likely to get good results by pushing AAI to provide better work than by switching contractors and incurring the learning costs with the new contractor. It appears that the surveys will be reasonably well administered. The major problems are to assure more continuous monitoring of the progress the programs and a deeper involvement of the analysts in the development of questionaires, specification of their analytical models and early analysis of baseline data. Steps have already been taken in

that direction by AAI, at the Foundation's insistance.

I believe the Foundation needs to appoint a person with professional experience in evaluation research to help it manage the research over the remaining course of the project. AAI complains about the lack of a consistent response from the Foundation on its inquiries and on delays in making key decisions as the research unfolds. The Foundation officers complain about slow and inadequate response from AAI. This is all normal for such projects, but I believe the perfomance on both sides would be improved if a good professional could be found to play the intermediary role for the Foundation. The major question is how much of an effort would it take and could an appropriate person be found an appointed.

I do not have time to pursue this question in detail here, but can do so later if the Foundation wishes. One important point is, however, that the person appointed should have some direct experience working for, or very closely with, a contract research firm. In order to obtain good performance from such firms, and to be reasonably responsive to their needs for direction, it is important to understand their internal dynamics.

In this appendix I will try to present some of the basic concepts relating to the use of statistical power calculations in the design of samples. I will use illustrative numbers that are reasonably close to those appropriate for MFSP.

Suppose a given type of experimental program was run not once but very many times, each time with a sample of 400 participants and 400 comparison group members. For each replication, average earnings for the comparison group were calculated and subtracted from the calculated average earnings of the participants. Call this difference in earnings the average program effect. The average program effect was recorded for each replication and a probability distribution was built up from the recorded program effects. Suppose that probability distribution looked like figure 1, having a mean of $600 and a standard deviation of $175. I will call this the distribution of sample average effects of the program and the mean of this distribution ($600 in this case) I will call the true average affect.

When an experimental program like MFSP is run, it provides just one sample drawn from the distribution of sample average effects. This observation is the key to understanding the use of statistical power in design of samples. The given distribution of sample average effects will give rise to a wide range of estimates of the average program effect, and when we draw a given sample for a project, we cannot be sure from which part of the distribution our particular sample will be drawn. We would like to design our sample so that we stand a good chance of not concluding that the true program effect is zero when, in fact, the true average effect is substantial. The statistical power of a sample is the estimate of the probability that a sample of a given size will not lead to the acceptance of the hypothesis that the true average efffect is zero when the distribution of sample average effects has a non-zero mean.

To clarify this further, it may be helpful to review how one will use the data collected from the sample to test a hypothesis about the magnitude of the program effect.

After the data for MFSP are collected, the estimate of the program effect (participant average earnings minus comparison group average earnings) is calculated, as well as an estimate of the standard deviation of the program effect. Using these data, an hypothesis test will be performed to determine whether the estimated program effect is statistically signficantly different from zero. This hypothesis test is necessary because even if the true average program effect were zero, samples would still occur which yield estimates of the program effect that differ from zero. The hypothesis that the true average effect is zero is often referred to as the null hypothesis. This hypothesis test may be visualized in terms of figure 2, which gives the probability distribution of sample average effects if the true average effect were zero and the standard deviation $175. The usual hypothesis test uses a level of significance of 5%. When this level of significance is used, the estmated program effect for the sample must be more than 1.96 standard deviations away from zero for one to conclude that the true average program effect is non-zero. Thus, in the example I have been

using, if the data yielded an estimated program effect of $400 (and the estimated standard deviation were $175), one could say that the program effect is different from zero at better than the 5% significance level (actually the 2.2% signficance level in this case); we would reject the null hypothesis of a zero true average program effect. (The 5% significance level means that we reject the null hypothesis with only a 5% chance of error in the sense that, if the true average program effect were zero only 5% of the samples of this size would yield estimates of the program effect that are more than 1.96 standard deviations away from zero; 5% represents the chance of Type I error, the error of rejecting the null hypothesis when it is true).

As noted above, the key point with respect to statistical power is that the sample result given by the MFSP sample is only one of many possible sample results that could be generated by the distribution of sample average effects. Another sample might yield an estimate of the program effect of $800, still another, an estimate of $300, and so on. Some of the samples would yield estimates of the program effect which would lead us to accept the hypothesis that the mean true effect is zero, even though the sample was generated by a distribution with the average true effect of $600. For example, this would be the case with the example distribution if the sample gave an estimate of the program effect of $300, since $300 is less than 1.96 standard deviations away from zero. When this occurs - accepting the zero effect hypothesis when it is false - it is called Type II error. A picture of the probability of Type II error that is likely if the distribution of sample average effects were as in figure 1, and a hypothesis test distribution as in figure 2 were used, is shown in figure 3. In that figure it can be seen that of all the samples that might be drawn from the distribution, a certain portion, shown as the shadded area, will yield estimates of the program effect that are less than 1.96 standard deviations away from zero and for these samples we would accept the hypothesis that the true average effect of the program is zero, i.e., we would commit a Type II error. For the distribution used in figures 1-3 the probability of Type II error would be about 7.2%, i.e., there would be a 7.2% chance of drawing a sample which would lead to acceptance of the hypothesis that the average true effect is zero when it is actually $600.

The power of the sample is the converse of the probability of Type II error; it is the probability that we will draw a sample that will lead us to reject the null hypothesis (of a zero true average effect in this case) when it is false. The power of the sample is illustrated in figure 4 by the cross-hatched part of the distribution; it is equal to one minus the probability of Type II error. In the example I have used, the power would be 92.8% (= 1-.072). To reiterate, if the distribution of sample average effects were as depicted in figures 1-3 there would be a 92.8% chance we would draw a sample which would yield an estimate of the true average effect which is statistically different from zero at the 5% signficance level.

For a given sample design, the power of the sample depends directly on the distribution of the sample average effects, i.e., its mean and standard deviation. For example, suppose the distribution of sample

average effects had a mean of $343 and a standard deviation of $175, as depicted in figure 5. It is easy to see that fully 50% of the samples drawn from that distribution would yield estimates of the average program effect which are less than 1.96 standard deviations away from zero. Thus, there would be a 50% chance of Type II error and only 50% power.

When we are designing a sample, we, of course do not know the location of the distribution of sample average effects (if we did, we would not need the sample to generate estimates). However, we usually have some information which allows us to make a good guess as to what the standard deviation of the outcome variable, e.g., earnings, might be for groups of the type of persons included in the program. (I have used AAI's estimate of the standard deviation of earnings for single female family heads for this purpose). With this guess as to standard deviation we can answer the questions: what would be the power of this sample if the true average effect were $200, if it were $300, if it were $600, etc.? In this fashion we derive estimates of the power of a given sample for various locations of the distribution of sample average effects. We can then say, for example, if the true average effect is $350, this sample of 400 participants and 400 comparison group members provides a 50% chance of obtaining an estimate of the program effect which will differ statistically significantly from zero at the 5% signficance level, i.e., the sample will have 50% power and if the true average effect is $600 this sample will have 92.8% power.

Suppose that we wanted to have a better than 50% chance of detecting (i.e. passing a 5% significance level test for a non-zero effect) a true average program effect as small as $350. Given the underlying estimates of the standard deviation of earnings used thus far, we would have to increase our sample size in order to achieve more than 50% power for a true effect that small. The effect of an increase in sample size is to reduce the standard deviation used for both the distribution of sample average effects and the hypothesis test. For example, if the sample size were increased to 820 participants and 820 comparison group members, the standard deviation of the distribution of sample average effects would be $122 and the power of the sample for an true average effect of $350 would rise to 80%. Figure 6 illustrates the power of this larger sample.

It can be seen from the above discussion, using the concept of power in sample design requires three major elements:

a. A good estimate (guess) of the standard deviation of the variable used as the outcome measure (an estimate particularly tailored to the type of persons whom the program is to enroll);

b. A choice of the level of power desired, e.g., 80% power (an 80% chance that the sample drawn will yield estimates that will pass a test for non-zero effects at the given significance level).

c. The size of true average effect for which the chosen level of power is to apply, e.g., we wish to have 80% power for an true average effect of $600.

With these elements given, one can determine the size of sample required.

It may be useful, as a sort of review, to consider how adding the concept of statistical power changes the sample design process from what it would be if one focused only on minimum detectable response. The minimum detectable response approach focuses on what one will do after the data are collected, ex post. It is reflected in the hypothesis test diagram represented in figure 2 where the minimum detectable response is that point 1.96 standard deviations away from zero (when the 5% significance level is used). The minimum detectable response focuses on Type I error.

For intelligent sample design, however, we need to try to form some a priori judgements. First, we must make a judgement about what level of program effect we would like to be able to detect, statistically, with our sample. Let us say, for example, we wish to be able to detect an effect as small as $350. It might seem that if we wish to be able to detect an effect as small as $350 that all we need is a sample size sufficiently large to have a minimum detectable response of $350, but this is not correct, because the true average program effect could be $350 but the sample we happen to draw yields an estimate of considerably less than $350. If this happened, our sample value would fall below the minimum detectable reponse and we would accept the hypothesis of a zero program effect;we would have a Type II error of accepting the zero effect hypothesis when it is false. How likely is this to happen? This is the second judgement we must form. We must guess what the likely distribution of sample average effects will be, that is, we must try to formulate our beliefs about about the shape and location of the distribution in figure 1. As soon as we realize that a given true program effect will give rise to a wide distribution of sample estimates of the average effect, we become aware of the risks of type II error and the need to plan a priori to reduce them to an acceptable level. That is, rather than focusing just on figure 2, as we do with minimum detectable response, we focus of figures 1 and 2 combined, as illustrated in figure 3 - 6. We want to set our sample size a priori so that we have a reasonable chance ex post of having a low level of both Type I and Type II error. So given our judgements about the level of effect we wish to detect and the likely location of the distribution of sample average effects, we choose the sample size which will give us sufficient chance to rejecting the hypothesis of zero program effect if the true average effect is at or above our chosen standard (e.g. $350). Our calculated chance of doing that is the power of the sample.

FIGURE 1

DISTRIBUTION OF SAMPLE AVERAGE EFFECTS
FOR SAMPLE 400 PARTICIPANTS (Np=400) 400 COMPARISON (Nc=400)
WITH TRUE AVERAGE EFFECT = $600 & STANDARD DEVIATION (σ) = $175

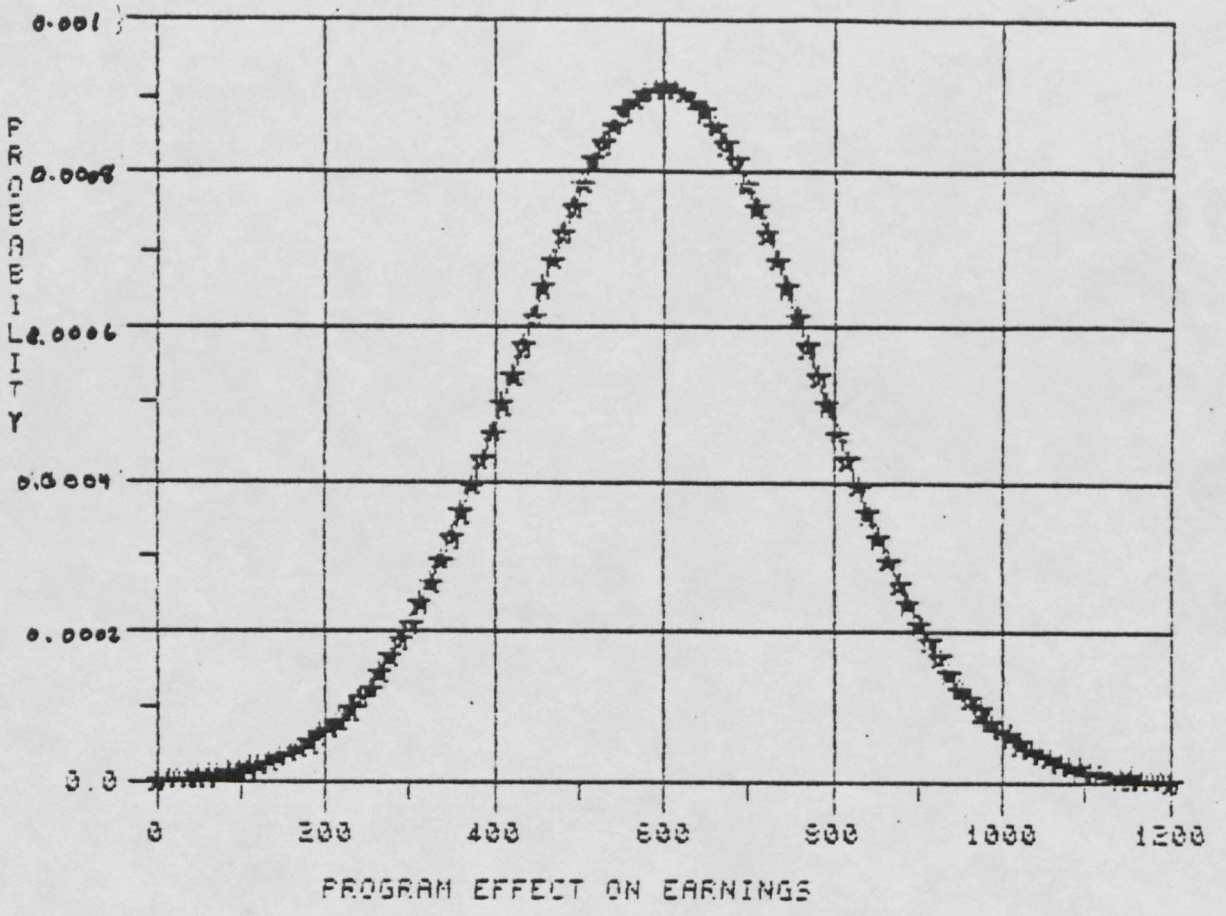


FIGURE 2

HYPOTHESIS DISTRIBUTION OF SAMPLE AVERAGE EFFECTS
SAMPLE Np=400, Nc=400
HYPOTHESIS TRUE EFFECT = 0 σ = $175

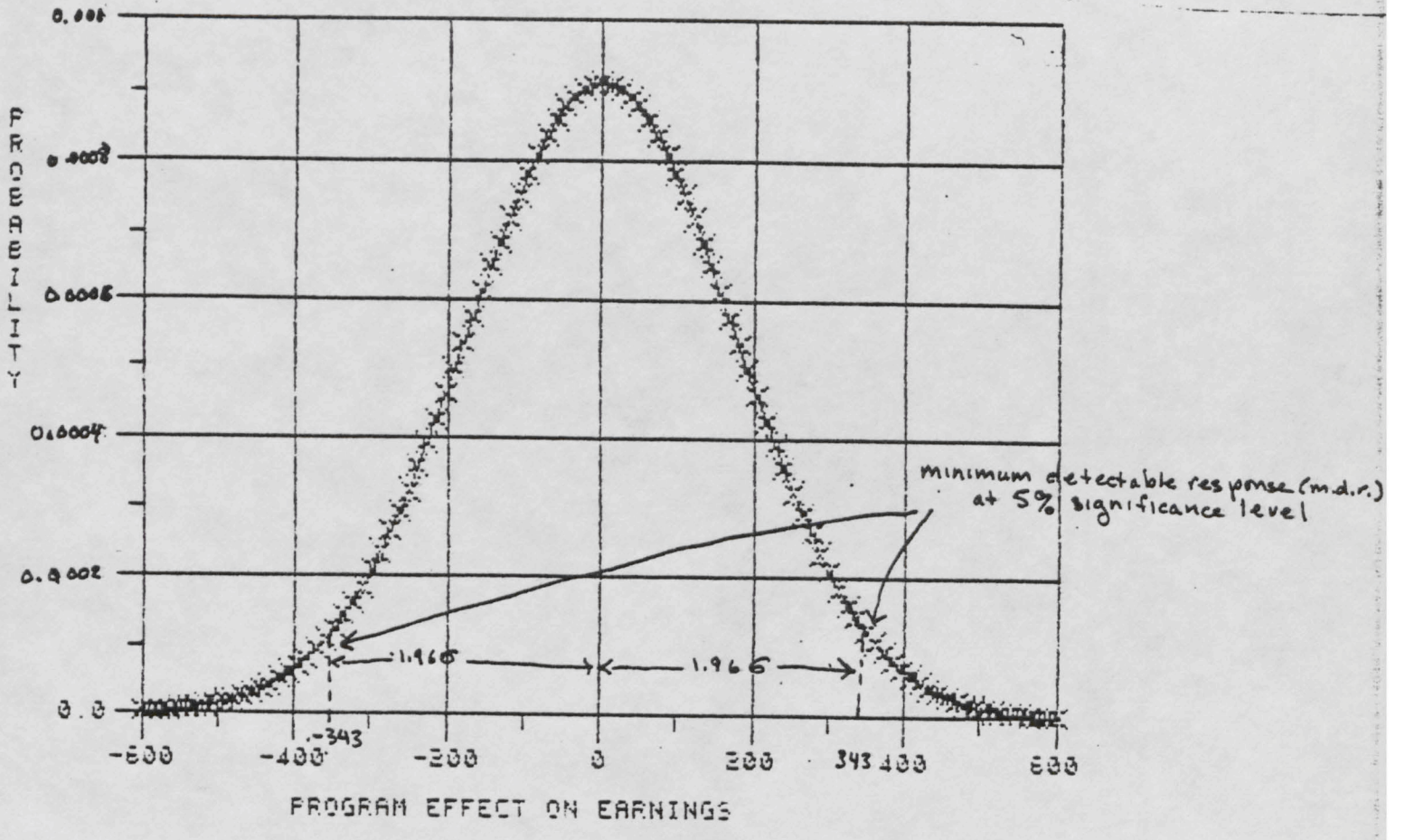


FIGURE 3

COMBINED HYPOTHESIS (FIGURE 2) AND TRUE SAMPLE DISTRIBUTION (FIGURE 1)

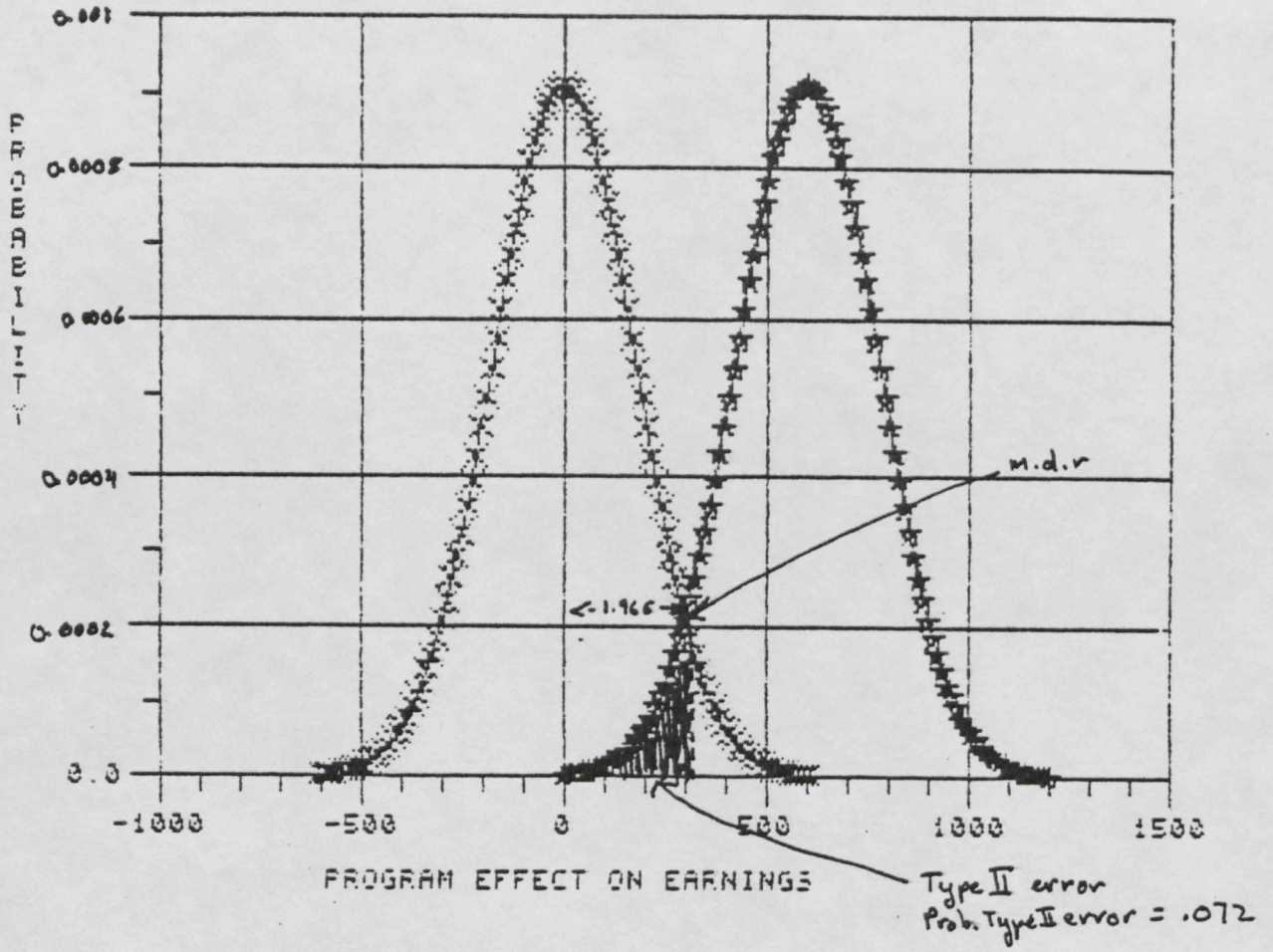


FIGURE 4

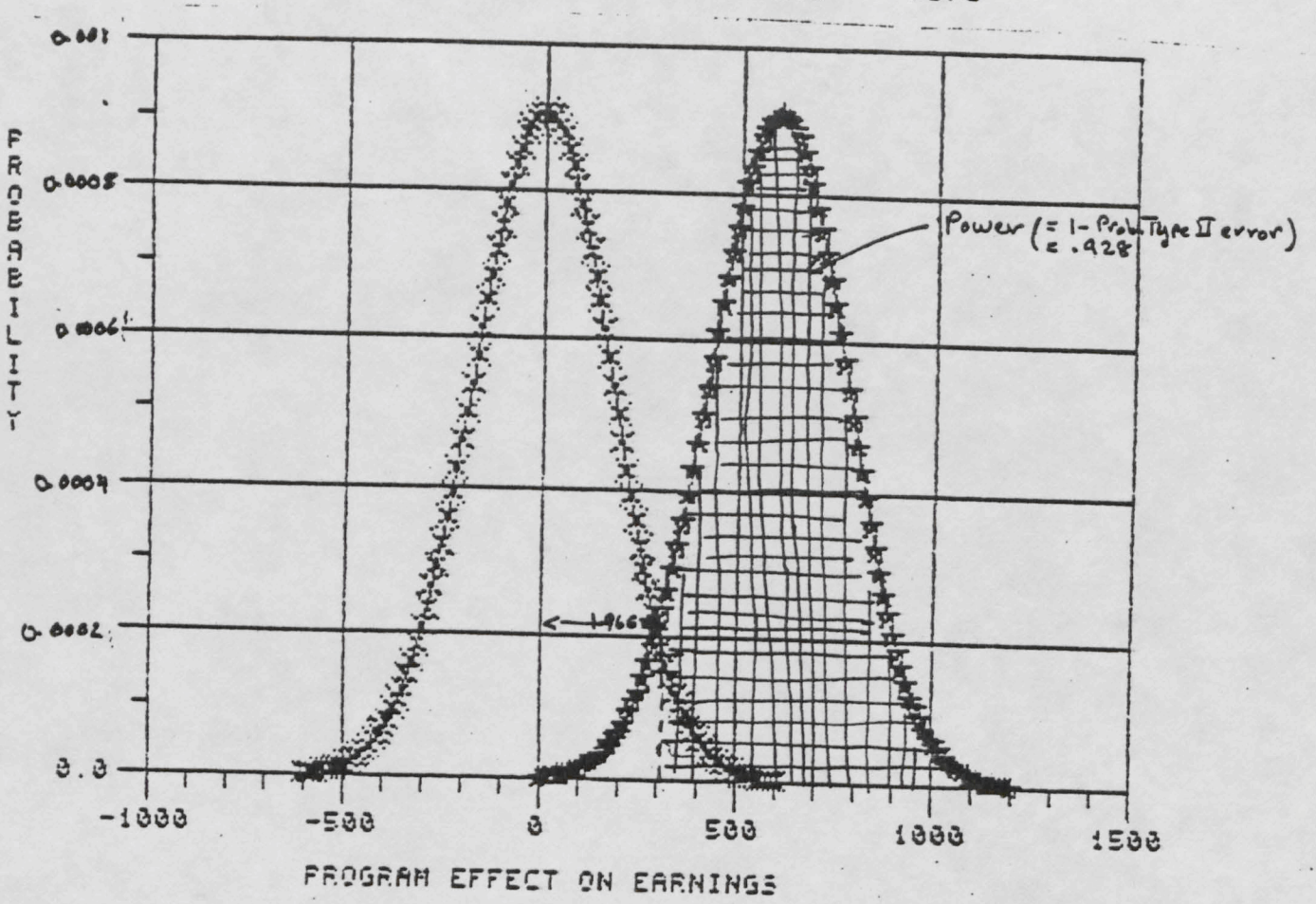

POWER OF SAMPLE
SAMPLE Np=400, Nc=400
TRUE AVERAGE EFFECT = $600 σ=$175
PROBABILITY
0.001
0.0008
0.0006
0.0004
0.0002
0.0
-1000
-500
0
500
1000
1500
PROGRAM EFFECT ON EARNINGS
Power (= 1- Prob Type II error)
= .928
<-1.96σ

FIGURE 5

DISTRIBUTIONS WITH TRUE AVERAGE EFFECT = $343

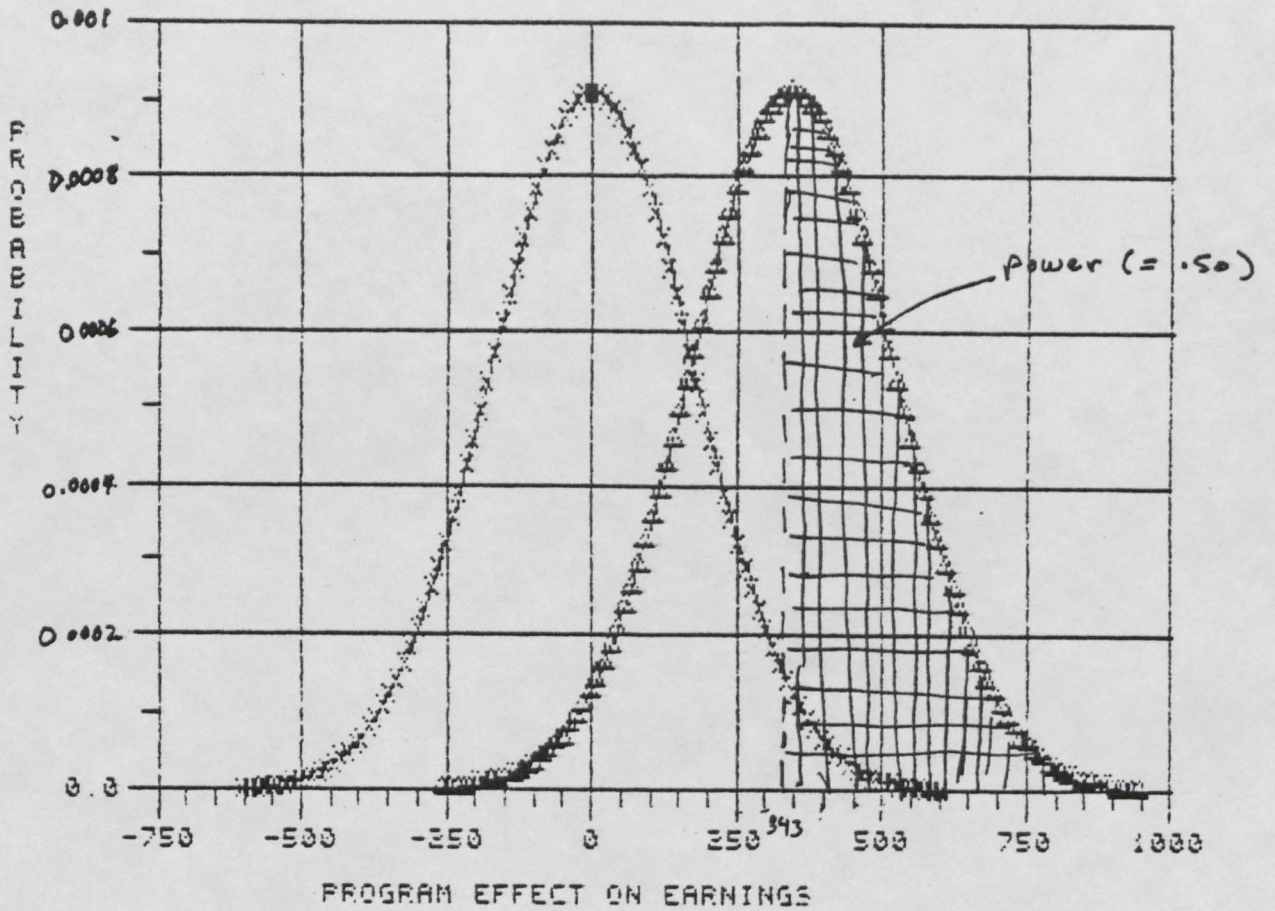
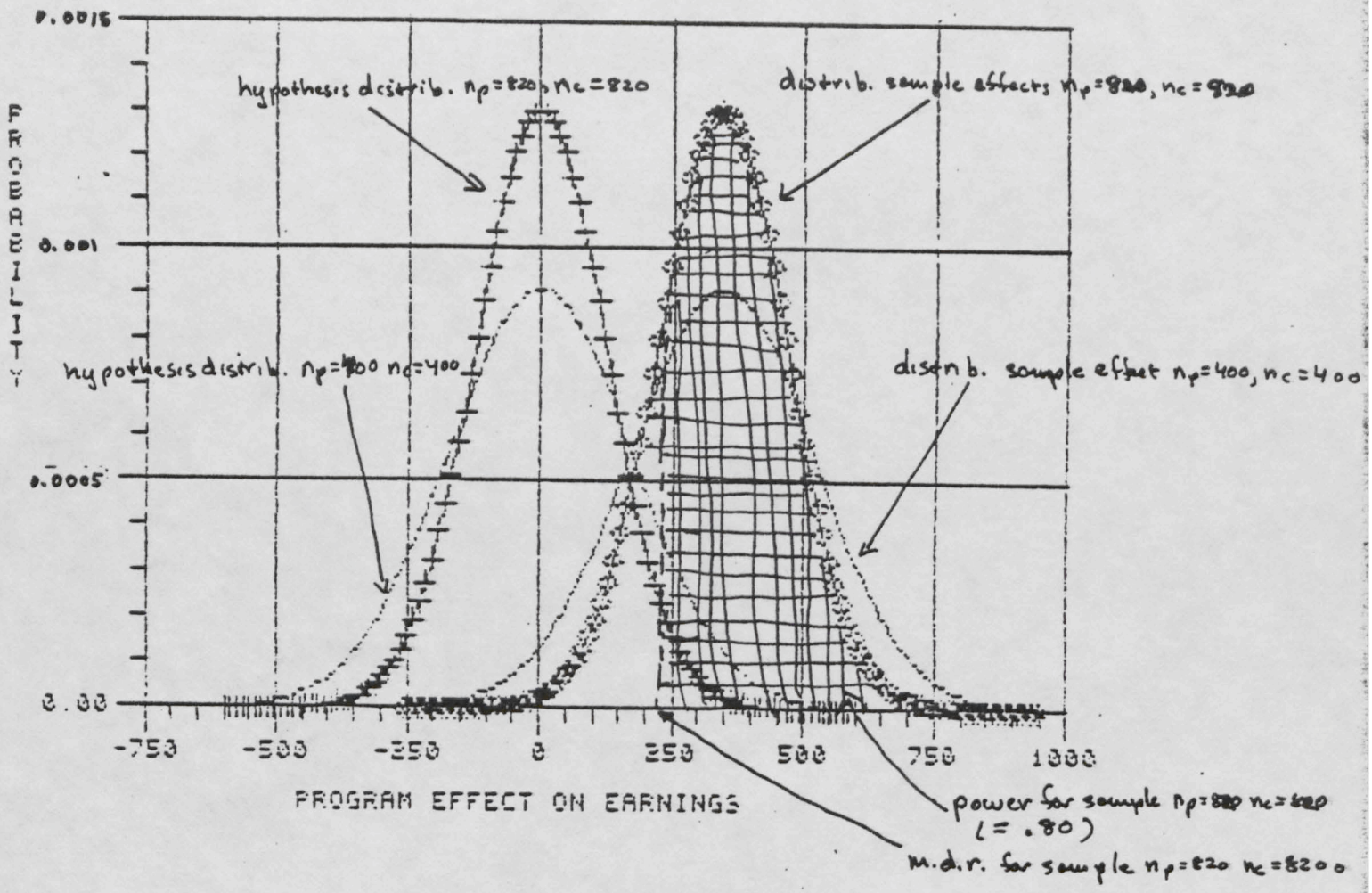


FIGURE 6

EFFECT ON DISTRIBUTIONS OF INCREASE IN SAMPLE TO Np=820, Nc=820



rer 7.6.84

To: Bernie Anderson, Phoebe Cottingham, Larry Stifel
From: Rob Hollister
Subject: New Estimates of Statistical Power

In my report dated 29 May 1984 I developed the estimates of statistical power for earnings using the estimate of the standard deviation of earnings developed by AAI. On page 10 of that report, I indicated by doubts about the appropriateness of their estimate of the standard deviation of earnings and urged that some further work be done to try to obtain better estimates.

I have recently obtained the estimates of the standard deviation of earnings from the Supported Work demonstration and I have done some quick calculations to see how they would alter the estimates of statstical power for the current sample design. Below I give you a table with revised estimates. Table I gives the estimates in the same form as in the report but using the revised standard deviation for earnings. Table II reproduces the 29 May table for comparative purposes. The estimates in Table I are just 1.33 times those in II, reflecting the larger standard deviation (and some more technical adjustments I go into below)

ESTIMATES OF STATISTICAL POWER
FOR THE SAMPLE DESIGN OF MARCH 27, 1984

TABLE I

SITE	MINIMUM DETECTABLE RESPONSE	TRUE EFFECT 90% POWER	TRUE EFFECT 80% POWER
	ANNUAL EARNINGS (IN DOLLARS) ASSUMING STANDARD DEVIATION = $3108		
ATLANTA	431	715	629
BROOKLYN	552	914	807
PROVIDENCE	463	767	676
SAN JOSE	450	745	656
WASHINGTON	490	811	715
POOLED	208	343	303

TABLE II

SITE	MINIMUM DETECTABLE RESPONSE	TRUE EFFECT 90% POWER	TRUE EFFECT 80% POWER
		ANNUAL EARNINGS (IN DOLLARS) ASSUMING STANDARD DEVIATION = $2453	
ATLANTA	324	537	473
BROOKLYN	415	687	606
PROVIDENCE	348	576	508
SAN JOSE	338	560	493
WASHINGTON	368	609	537
POOLED	156	258	228

The estimates in Table I indicate that the project is much closer to the margin of adequate sample size for detecting effects on earnings that had been indicated in the estimates provided in the 29 May report (and in Table II). Whereas the earlier estimates had indicated that if the true average effect were $600, the sample would provide 80% power or better for virtually all the sites, the new estimates indicate only if the true average effect were about $700 would the sample power be about 80% for the sites, and for Brooklyn, 80% power would only be obtained for a true average effect of $800.

Phoebe asked me to also calculate, using the new estimate of the standard deviation, how much of an increase in sample would be necessary to return the power to the level given in Table II, i.e., how much bigger samples would be necessary to obtain the level of statistical power indicated in the 29 May report. The answer is that all sample sizes would have to be increased by 77%, e.g., for Atlanta the planned number of followup interviews was 364 for participants and 371 for comparison group but they would have to be increase to 645 for participants and 657 for comparison group in order to have a minimum detectable response of $324, 90% power for a true average effect of $537 and 80% power for a true average effect of $473.

For the purposes of the record, I report a few technical details regarding the new estimates in relation to the old.

For Supported Work, the estimated standard deviation in earnings for the AFDC group at the 18 month interview was $3108. In the regessions used to estimate the program effects, the regression adjusted standard deviation of earnings was $2981. In AAI's estimates they used a standard deviation of $2453 and assumed the regression adjustments would reduce this to $2124 (this is their assumption that the regression would have an R of .25). In addition, AAI makes an allowance for the correlation of comparison group status with the other

demographic variables used in the regression to estimate the program effects (this is their Rx.t of .10). Since Supported Work had random assignment, there was no correlation between the control group status and the other demographic variables. To adjust the estimates in Table II for both these factors I increase the estimated deviation but remove the correction for correlation with of comparison status with other variables. Thus to derive the new estimates I use the following correction:

$$\sqrt{\frac{(2980)^2}{(2124)^2} * 0.9} = 1.33$$

One could argue about the appropriateness of using the Supported Work figures in making these estimates. I will not detail the possible arguments and responses here, but will be glad to do so if anyone wants to review them. The major point is that the Supported Work estimates are derived from a program that covered a roughly similar group and appear to give more conservative estimates of statistical power than AAI's alternative. The burden of the argument would seem to fall on those who would contend that the Supported Work estimates are too conservative.