POTENTIAL FOR PLANNED EXPERIMENTATION IN THE DEPARTMENT OF LABOR REGULATORY AREA

Stanley Masters; David Zimmerman; Karen Holden; James Jones, Jr.; Richard Kaluzny; Susan Meives; Craig Olson

SR23

This report was prepared by the Industrial Relations Research Institute, University of Wisconsin-Madison, and Mathematica Policy Research, Inc., for the Office of the Assistant Secretary for Policy, Evaluation and Research, U.S. Department of Labor, under contract/purchase order No. J-9-M-7-0153. Since contractors conducting research and development projects under Government sponsorship are encouraged to express their own judgment freely, this report does not necessarily represent the official opinion or policy of the Department of Labor. The contractor is solely responsible for the contents of this report.

PREFACE

We undertook this study primarily because of our interest in experimentation research and Department of Labor (DOL) regulatory programs, and our belief that more research was needed on the effects of and possible improvements in these programs. Gerald Somers was the main force in developing initial ideas for the study, and in assembling a research group with similar interests and particular areas of expertise in the regulatory areas. In particular, Jerry stressed the greater feasibility of varying enforcement strategies as the experimental treatment rather than varying the regulations. He also emphasized the need to establish good relations with government officials in the offices of the Department of Labor that were under study. Jerry died very unexpectedly at the end of December 1977, just as the main portion of the study was to commence. His death was a great personal loss to each of us.

With encouragement from ASPER, we decided to continue with the study, limiting the focus to OSHA, ERISA, and OFCCP. The early months of the project were spent familiarizing ourselves in more detail with the legislation in these areas, developing contacts in the DOL offices charged with the regulatory and enforcement functions, and generating a preliminary list of potential research topics in each of the areas. The next major activity was a conference in March, which included officials from ASPER and the programs under study, researchers interested in these programs, and representatives of firms and unions. Preparing for this conference took considerable effort

in developing the general experimentation ideas we had had at the start into explicit and precise statements about possible strategies so that we could receive useful feedback from conference participants. In our view the conference was very successful, both in terms of generating new ideas for experimental and nonexperimental research topics and in eliciting useful reactions to the ideas already developed. After the conference we wrote up summaries of the discussion and developed preliminary designs for a number of experiments based on these conference discussions. Early in July we circulated an interim report that included these summaries and preliminary designs.

We had hoped to receive reactions to the interim report that would enable us to limit our focus to a smaller number of specific experiments, so that we could develop more detailed designs and talk with agency, management, and union officials about implementation issues. Although we did receive helpful comments on this report, no consensus developed on which of the possible experiment most warrant further attention. In the absence of such a consensus, we decided to place more emphasis on general design considerations and legal issues rather than focus on specific experimental possibilities that might not conform to agency priorities.

We believe we have been successful in identifying both a useful set of possible experiments in the areas of OSHA, ERISA, and OFCCP, and in outlining the major design issues that must be considered in future experimentation activities. But any such activities must begin with decisions by the Department of Labor on which experiment(s) would be most useful and policy-relevant. Further efforts could then be undertaken to design and implement pilot versions.

ABSTRACT

The objective of this study was to provide the Department of Labor with information on the feasibility of conducting experiments to assess the effects of possible changes in three of its regulatory programs--the Occupational Safety and Health Administration (OSHA), the Employee Retirement Income Security Act (ERISA), and the Office of Federal Contract Compliance Programs (OFCCP). We focused on two important issues: (1) the identification of specific policy questions with regard to OSHA, ERISA, and OFCCP that can be addressed with experimental research and that may be of sufficient importance to warrant undertaking such research; and (2) an examination of important design issues that need to be addressed if any of those experiments are to be undertaken, including the specification of experimental treatments and outcomes, the duration of the experiment, the unit of analysis, and the prospects for cooperation or noncooperation by affected firms and workers, which also covers the possibility of legal challenges. Our findings are based on a review of the literature and discussions with government officials, labor and management representatives, and many leading policy researchers.

We found that there is considerable interest in experimentation with regard to the three regulatory programs--spread over many possible topics, however. We concluded that the most appealing candidate for experimentation in OSHA is variation in targeting strategies, and that other possibilities include varying the average probability of inspection and/or reinspection, and providing incentives for the formation of

effective labor-management committees on workplace safety and health. For ERISA, the most promising candidates are variations in what plan administrators are required to report to the government, what they must disclose to enrollees, and variations in Pension Benefit Guarantee Corporation (PBGC) premiums. In OFCCP, the best candidates are variations in targeting of complaince reviews, possible financial incentives for government contractors who have good Equal Employment Opportunity (EEO) records, and possible training subsidies for those with weak EEO records.

If the idea of possible experiments with regard to its regulatory programs continues to be of interest to DOL, it must decide which issue or issues are most appropriate for experimental research. For each experimental possibility under active consideration, the set of experimental design issues must be addressed. Next, a small pilot experiment should be developed. Only if each of these preliminary activities are successfully completed can a full-scale experiment be undertaken with any reasonable prospect for success.

TABLE OF CONTENTS

continued

TABLE OF CONTENTS (continued)

LIST OF ILLUSTRATIONS

For any experimental study to make a valuable contribution to policy considerations, it is important that the topic be well chosen, that the dependent variables (effects of the policy alternatives) be well thought out and carefully measured, and that there be enough variation in the experimental treatment (the policy alternatives being examined) so that, for the sample size available, there is a reasonable chance of detecting the effects of differences in the treatment. In addition, the experimental treatment must not covary so closely with any other factor (or set of factors) that it is impossible to separate the effects of the experimental treatment from that of the contaminating influence.

In a classical experiment, the possibility of such contamination is greatly reduced by random assignment to various treatments. In a natural or quasi-experimental context, in contrast, where the analysts have no control over who is assigned to the various treatments, the possibilities of contamination are more severe. For example, if the policy variable is whether an establishment was inspected by OSHA during a given year, and if the dependent variable is the establishment's accident rate during that year, it is quite likely that the effect of OSHA inspections on accident rates will be confounded by the fact that OSHA inspections are more likely in establishments with high accident rates.

It is possible to overcome the contamination of results caused by this "selection bias" problem if the firms selected for inspections (or chosen for some other experimental treatment) are picked by a formula and if this formula is known by the researchers. However, if inspection probabilities are based in part on subjective factors, such as whether a complaint appears to be legitimate and serious, then a more complicated approach is

the potential for experimentation and discusses the next steps that should be taken if DOL continues to be interested in experimental research in these areas.

Although identifying important policy related research questions is an obvious prerequisite to considering the desirability of experiments to answer such questions, it is not obvious that the experimental methodology is the most appropriate method for addressing many of them. Thus, it is important to briefly contrast the classical experimental methodology with the "natural" or quasi-experiment that is often the basis for policy evaluation studies.

A policy experiment, whether classical or "natural," is the introduction of a potential government policy (or a change in an existing policy) on a limited scale with carefully planned data collection and analysis efforts in order to help government decisionmakers assess the probable impacts of larger scale implementation of that policy.[1] In this study our focus has been on classical experiments, in which the subjects (in this case employers or groups of employers) are randomly assigned to different treatment alternatives that are specified as part of the experiment. In other words, the experimenters have control over both the treatment and the assignment to those treatments. In contrast, in a natural or quasi-experiment the researcher often has no control over the policy specification, the manner in which it is administered, or the sample to which it is administered. Consequently, the conclusions that can be drawn and the degree to which they can be generalized are more limited.

---

[1] This definition is adopted from Kaluzny and Ohls (1976).

POTENTIAL FOR PLANNED EXPERIMENTATION IN THE DEPARTMENT OF LABOR REGULATORY AREA

1. INTRODUCTION

The goal of this study was to determine the feasibility of conducting experiments to assess the effects of possible changes in the regulations and enforcement of three Department of Labor (DOL) regulatory programs: the Occupational Safety and Health Administration (OSHA), the Employee Retirement Income Security Act (ERISA), and the Office of Federal Contract Compliance Program (OFCCP). Although some of the controversy surrounding these programs has focused on program goals, a large amount of the ongoing dispute has been centered on how effective these programs have been in achieving their objectives and on the costs associated with the gains that have been achieved. Much can be learned about issues of cost effectiveness from experimentation, including the potential cost effectiveness of alternatives or supplements to existing policies.

Given this context, the study has focused on identifying specific topics that might be most appropriate for experimentation within each of the three regulatory programs. In Section 2 of this report we address a series of issues that ultimately shape the design of any potential experiment: the specifications of possible treatment variables, the identification of operational measures of treatment effects, the establishment of an appropriate experimental time period and unit of analysis, and the problems of enlisting and maintaining the cooperation of participating firms. Sections 3, 4, and 5 of the report discuss in more detail the design of potential experiments in each of the three regulatory areas. Section 6 summarizes

necessary, based on stronger assumptions concerning the selection process.[2] Although this approach appears to be promising and warrants further work, it is difficult for the noneconometrician to understand just what is being done and, in particular, to assess the plausibility of the assumptions being made.

Recent policy evaluation studies by Heckman and Wolpin (1976) on the Contract Compliance Program and Smith (forthcoming) on OSHA show that it is possible to deal with the selection bias problem by making plausible assumptions about how the selection process is likely to be related to the outcome measure and then using these assumptions to control for the effects of the selection process. Although these studies both show considerable imagination, neither can solve all the problems that occur in the absence of the two essentials of a classical experiment: random assignment, and research control over the experimental treatment (see Appendix A for a critique of these studies). Thus, there still appears to be a strong case for using random assignment wherever possible in evaluating the effect of policy alternatives.

The classical experiment is a much more flexible tool for policy analysis than the typical evaluation study. In part, this is a result of the focus on several alternative treatments or treatment levels within the context of the experiment. In part, it is because the design exercise forces the researchers to consider and specify the exact nature of the experimental treatment. Moreover, the design of an experiment requires proponents of proposed policy changes to think through all the details necessary in order

---

[2]For a good discussion of various ways of dealing with the selection bias issue, see Barnow et al. (1978). For an application of the new approach, see Katz (1978).

to implement their policy.[3] In addition to being a valuable exercise in and of itself, the resulting knowledge of what is being evaluated is an important advantage of experiments. As a result, the experiment can examine the components of any given policy and draw conclusions about the effect variations in these components will have. This increases the usefulness of the results as guidelines for designing policy alternatives because the effects of policies which involve these components but which were not explicitly tested in the experiment can be predicted on the basis of the experimental response.

In contrast, the methodology for evaluating a current policy provides some idea about the effectiveness of that particular policy but offers little insight into how modifications of that policy might improve its effectiveness. In addition, the evaluations of current policy often tend to focus disproportionately on the problems of measuring effects without worrying about specifying the treatment stimulus. For example, OFCCP studies have normally paid little attention to what is actually done during a compliance review.

Despite their advantages, large-scale classical experiments on policy alternatives have only recently been undertaken. The primary obstacle appears to have been the view that it is unfair to subject otherwise equivalent individuals or institutions to different treatments simply to facilitate research. Until the New Jersey Negative Income Tax Experiment, which began in 1968, no such experiments had been undertaken in this country. Once this experiment was successfully undertaken, however, others rapidly

---

[3] For example, in the negative income tax (NIT) experiments, researchers advocating NITs were forced to decide issues such as the appropriate accounting period, the definition of the recipient unit, and the way in which the NIT should be integrated with other transfer programs.

followed--including income-maintenance experiments at other locations, and experiments on health insurance, housing allowances, and the Supported Work Demonstration.[4]

Although no prohibitive legal or political problems arose in these experiments, it should be noted that the subjects were all relatively disadvantaged individuals or families whose opportunities were being improved by the experiment. For the DOL regulatory programs, however, the subjects would be firms (or groups of firms). If a firm believes it is being hurt by an experiment (e.g., its costs increased or its position weakened relative to some of its competitors), then the firm is likely to institute a more vigorous legal and/or political challenge than might be expected from disadvantaged clients whose economic position is being improved (held constant, for controls). Although it is not clear whether any legal or political challenges would be successful, the above argument does suggest the need to plan any experiment so as to minimize the chances that it would hurt the condition of any participant.

The legal issues, which we have discussed with the Solicitor of the Department of Labor, Carin Clauss, do not seem to be very clear cut. For example, there is a tension between the policy value of the knowledge gained from experiments and the "unequal treatment of equals" inherent in any research design based on random assignment to experimental or control status. The tension has a quantitative as well as qualitative dimension, since a reasonably strong experimental treatment greatly increases the chances of being able to detect experimental effects, but the stronger the treatment

---

[4] Supported Work is a work experience program that utilizes random assignment to experimental and control groups, although the experimental treatment is not under the direct control of the researchers.

the greater the unequal treatment of experimentals and controls and the greater the potential for litigation. Given the possibility of legal challenges (and their potentially disruptive influence even if they are not ultimately successful), governmental officials are generally reluctant to undertake experiments in the absence of specific authorization in the legislation (or executive order in the case of OFCCP). The OSHAct does in fact provide that the Secretary of Labor (or HEW) can grant variances for experimental purposes.[5] Such explicit authority is not, however, provided in either ERISA or the executive orders establishing OFCCP.

Despite the legal uncertainties, we believe that experimental possibilities should be actively considered. Since some legislation (including the OSHAct) does explicitly authorize experiments, presumably other legislation (such as ERISA) and executive orders could be amended to include a similar authorization.

This report focuses primarily on identifying important topics that might be appropriate for experimentation within each of the three programs. We chose to emphasize possible variations in enforcement rather than on changes in the regulations themselves, since DOL is expected to have greater flexibility with regard to enforcement procedures. Although the emphasis is on potential classical experiments, much of the discussion also relates to evaluations of natural experiments.

---

[5] Section 6, paragraph 6C, of the OSHAct reads as follows:

> The Secretary is authorized to grant a variance from any standard or portion thereof whenever he determines, or the Secretary of Health, Education, and Welfare certifies, that such variance is necessary to permit an employer to participate in an experiment approved by him or the Secretary of Health, Education, and Welfare designed to demonstrate or validate new and improved techniques to safeguard the health or safety of workers.

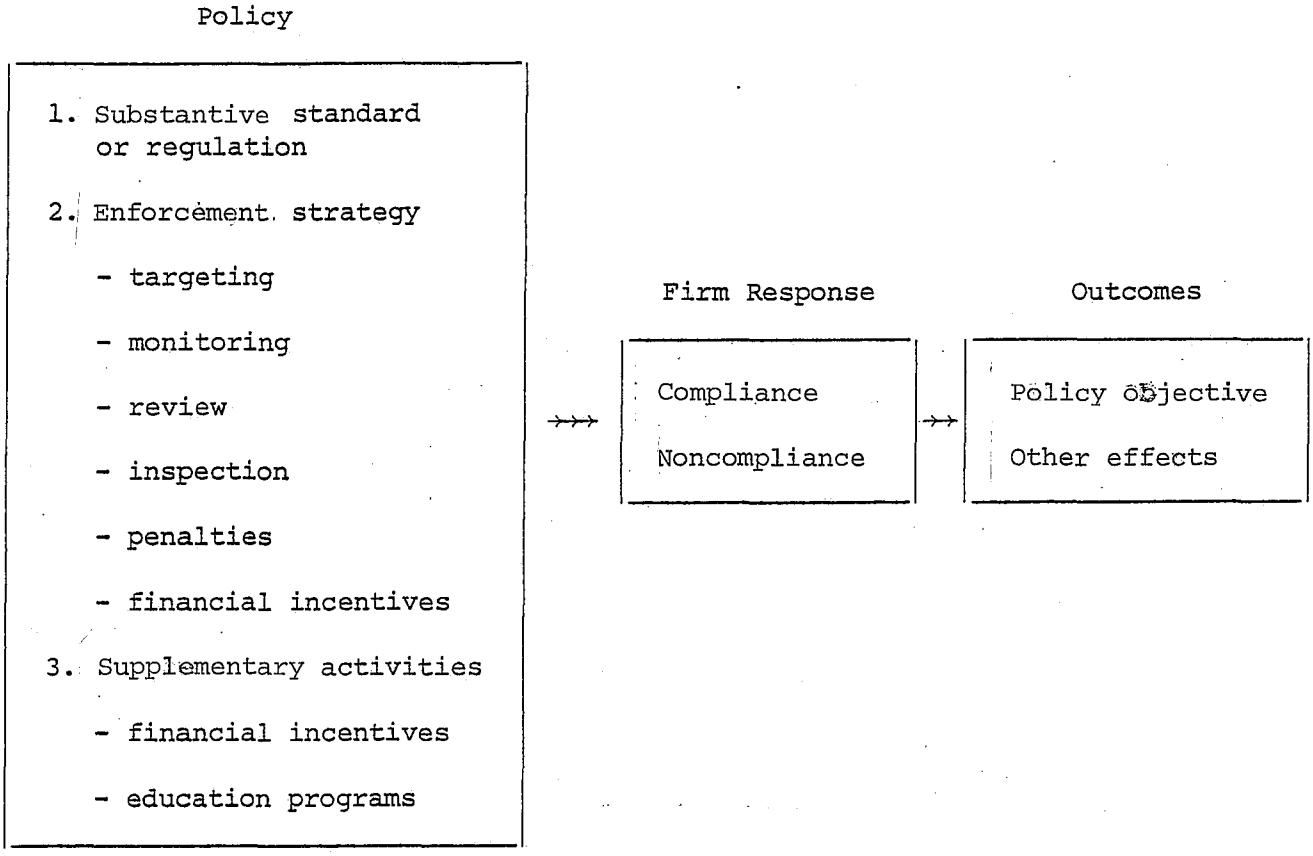
2. EXPERIMENTAL DESIGN ISSUES

Experimental Treatment

Although the substantive content of regulatory policies varies considerably across the three program areas, there are many functional similarities among them. These become more apparent if we represent the regulatory process as a simple sequence of events (Figure 1).

Figure 1

Policy Elements in the Regulatory Process



The regulation per se is only one of three components that make up the total regulatory policy. The second component is enforcement, which represents elements such as the level of inspection and review devoted to achieving compliance with the regulations, the magnitude and nature of penalties, and the method of allocating agency resources among enforcement activities. The third component is other activities outside the substantive regulation or the actual enforcement process, such as either the provision of financial incentives for compliance or education programs, both of which can be important elements in regulatory policy. Experimental treatments (i.e., variations in the regulatory policy) can be specified in terms of any of the three components. In some areas, like ERISA, the substantive policy includes reporting or disclosure requirements, with enforcement being primarily a monitoring function to ensure the requirement was fulfilled. In areas like OSHA or OFCCP, the enforcement strategies are an integral part of the regulatory policy because the effective level of compliance is jointly determined by the stringency of the policy standard and the extent of enforcement.

In our discussions with agency staff members, researchers, and management and union representatives, we have found particular interest in enforcement strategies, reporting requirements, and financial incentives. The specification of potential experimental treatment variables in these three areas is discussed in detail below.

Enforcement strategies. The analysis of enforcement activity in an experimental context is appealing on two major grounds. First, unlike changes in the substantive policy standard, which may be subject to legal and political constraints, enforcement activity is primarily determined within the agency or program. Second, because resorces are

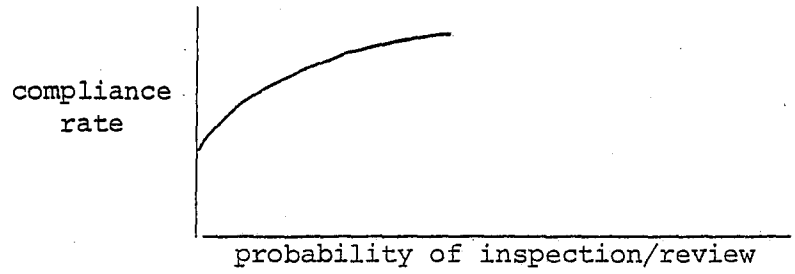
limited, current enforcement practices are not universal but fall instead on some subsample of firms. That is, within the current policy for each agency there is already variation in enforcement levels, although that variation tends not to be very systematic or readily analyzed.

The research issues involved in evaluating enforcement strategies can be expressed in terms of Figure 2, which shows that for any given level of penalty we can expect a positive relationship between enforcement activity and compliance--as the probability of inspection/review increases, the rate of compliance increases. Given this basic relationship, two issues are subject to experimental testing and verification:

1. By how much does increasing the level of inspection/review increase the extent of compliance?
2. What is the most productive way of allocating a given amount of enforcement resources among firms?

Figure 2

Relationship Between Probability of Inspection and Compliance Rate



The first issue amounts to estimating the shape of the compliance-enforcement relationship for the average firm in a given industry. Ideally,

if the level of penalty (e.g., fines) could vary as well as the level of inspection, a series of compliance-enforcement relationships could be examined. We would expect, for example, that as penalties increased the relationship would pivot upward to the left (i.e., enforcement activity would become more effective as the cost of noncompliance increased).

In an experimental context, the probability that a firm will be inspected can be interpreted as the treatment. For example, in a very simple experimental design, suppose that the probability of inspection in any given period can have three values: $T_1 = .10$; $T_2 = .50$; $T_3 = 1.0$. Then, if a sample of firms were randomly assigned to the various treatment levels, it would be possible to estimate the positive relationship between enforcement and compliance rates.

This simple framework can readily be extended to address the second issue--targeting of enforcement resources. Operationally, targeting means that certain types of firms (i.e., those satisfying the targeting criteria) face a considerably higher probability of inspection than the average firm.[6] Basing the targeting on the performance of the individual firm (perhaps relative to other similar firms), for example, will reduce the burden on firms with good performance records and will give firms a financial incentive to perform better--assuming that greater probabilities of inspection and greater reporting requirements are expensive for the firm and that the firm is aware of the targeting strategy in use.

---

[6] Some targeting currently exists but, aside from complaints that it may or may not be highly correlated with actual compliance, the targeting is based only on contract size (OFCCP) or mainly on industry performance (OSHA).

To make the targeting formula the treatment, targeting formulas can be developed a priori, with different ones randomly allocated for use with various subsamples of firms. Inspection probabilities for individual firms within a given group are then determined by the treatment formula assigned to the group. Effects on compliance or accident rates can then be observed for each group as a whole. (Further consideration of such an approach in OSHA is discussed in Section 3.)

Although the discussion so far has presented a treatment variable as a simple single-dimension concept; in fact, experimental treatments can be viewed as treatment packages defined by several components. For example, two primary factors in an enforcement strategy are the average probability of inspection and the extent of targeting. Although experiments could be designed using treatment variables defined in terms of either the level of inspection or the targeting strategy used, a more efficient design would combine both design treatment elements into a single experimental context.[7] The array of treatment possibilities can be represented as follows.

Figure 3

Experimental Treatments

	Targeting			
Inspection Frequency	None	T_1	T_2	T_3
low				
high				

[7] The experimental design does not necessarily have to include all possible treatment combinations. The most efficient design may actually exclude the combination most likely to be adopted (see Metcalf, 1977). It is necessary only that combinations be included that allow estimation of the way in which treatment effects vary.

These combinations of targeting formulas and probabilities of inspection define potential treatments. Additional treatments can be defined, obviously, by specifying exact inspection probabilities, additional targeting formulas, or adding reinspection probabilities. One advantage of combining these factors into a single design is to emphasize the tradeoffs involved in the enforcement process. Several interesting research hypotheses and policy questions are readily apparent.

1. How effective are alternative algorithms in reducing hazards?
2. Is targeting as effective as increasing the probability of reinspection?
3. Are strategies of targeting or increasing reinspection probabilities as effective at high levels of initial inspections as they are at low levels?

Reporting requirements. One type of policy standard which may be subject to relatively easy variation within each program is the information reporting requirements. These could include both requirements to report to the agency as well as requirements to report or disclose information to employees. The former is most applicable in ERISA and OFCCP when the employer must prove compliance. The latter is most applicable in OSHA and ERISA where one method for achieving program objectives is to provide employees with information that increases awareness of dangers or rights.

Several factors could be varied to define alternate treatments. One complaint that is common to both employers and agencies is the cost and difficulty of keeping up with the volume of required reporting. In an experimental context, both the type of information required and the frequency of reporting could be varied. Complex treatments could be defined, as in

the case of enforcement strategies, which combine variations in the type of data required and the frequency of reporting with variations in requirements by type of firm (i.e., targeting the reporting requirements).

Financial incentives. The converse of penalties and fines in the enforcement of regulatory standards is a set of financial incentives. In both cases the firms are faced with an increased cost of noncompliance, either directly as a penalty or indirectly as a subsidy or cost saving foregone.

Offering direct, obvious financial incentives to firms to increase compliance is a topic that has come up in both OSHA and OFCCP. Politically, the opposition to this approach appears quite strong in the case of OSHA, but not necessarily in the case of OFCCP--perhaps because there has been less prominent advocacy of the financial incentives approach for OFCCP. Because of this and also because financial incentives and penalties might be especially powerful techniques in OFCCP enforcement strategies--where they might gain added leverage by affecting the eligibility or ability of the firm to pursue certain types of contract--we have given the most attention to possible experimentation with financial incentives in our OFCCP discussion.[8]

Outcome Measures

The specification of measures by which an experimental policy is to be evaluated poses both conceptual and operational difficulties. Three issues must be addressed: (1) identification of the objectives of the program;

[8] Incentives in the form of cost savings generated by exemptions from certain reporting requirements may have weak effects but they may be more generally applicable and acceptable than direct financial incentives.

(2) specification of other effects (variously termed side effects, secondary effects, or unintended consequences) of the program which may be important; and (3) development of operational measures of experimental effects.

Specification of objectives. The authorizing legislation usually sets forth the goals of the program in general terms. These goals must be clarified and defined more precisely, however, if they are to serve as the basis for defining and measuring experimental outcomes. The process of clarifying objectives has two parts: (1) separating objectives into single-dimension elements; and (2) specifying standards of performance. Once these steps are completed, the researcher has a better idea of which objectives are actually addressed by the experimental treatment and at least a preliminary idea about the data required to measure the effects of the treatment.

For example, the goal of ERISA is to reduce the risk of workers not receiving adequate pension benefits despite long-term participation in a firm's pension plan. It seeks to reduce this risk by provisions designed to increase both the adequacy of the organization's pension fund and the probability that employees will be eligible to draw from that pension fund upon retirement. Although they are both equally important, the objectives of increasing fund adequacy and worker eligibility are quite different from each other, as are the regulations and policies designed to achieve them.

The disaggregation of regulatory program goals into more specific objectives is useful in the other two regulatory areas as well. For example, separating the health and safety objectives of OSHA is important because they present very different measurement and research design problems. Similarly,

the OFCCP goal of increasing the employment opportunities available to minorities, women, and other groups can be disaggregated into the objectives of increasing the proportion of minorities, women, etc. who apply for employment and then increasing the proportion of these applicants who are selected.

The second step, specification of standards of performance, provides a means of highlighting any complexities and special conditions of the policy to be evaluated that may be important considerations in designing an evaluation framework. In some cases, performance standards (e.g., employee eligibility) are relatively easy to identify. In other cases, where the objective is stated in more subjective or abstract terms (e.g., fund adequacy), the performance criteria are less apparent. In both instances, issues concerning the population to which the measures should apply (i.e., do the provisions of the policy exempt certain types of workers or do special circumstances affect more workers in one industry than in another) and the time period over which they apply (i.e., is adequacy defined relative to current status, expected short- or long-run needs, or under a range of assumptions about future needs) must be addressed.

Specifying secondary effects. Outcome measures defined in terms of one or more program objectives are necessary elements in determining whether an experimental policy works as planned. They do not, however, provide sufficient information to determine whether the experimental policy is feasible or desirable. For this, possible secondary effects must also be assessed.

The importance of secondary effects is underscored by the lack, in many cases, of a single dimension response to the treatment application.

This is especially likely in the case of enforcement strategies, where the response by the firms may range from noncompliance to several different technological methods of compliance. The treatment effect, then, may be reflected in measures of program objectives (e.g., accident rates) as well as in other characteristics of the production process, such as the size and composition of the labor force, the product price, relative wage levels, and the induced demand for new services and/or products.

Development of operational measures. Even if the objectives of the program and other likely effects of the treatment variables can be precisely defined, they must be correctly measured in order for an experiment to assess the effectiveness of changes in the program. As noted previously, specification of standards of performance in attaining objectives may not always be easy, especially when objectives involve subjective concepts such as adequacy of pension funds or availability of minority workers. The process of specifying conceptual measures of desired program outcomes requires that subjective choices be made regarding the relative importance of alternative aspects of the definition. The process is, however, further complicated by consideration of operational features of those specifications.

One major problem (discussed in the next division) is that the desired policy effects either may not occur or may not be measurable within the observation period of the experiment. A second major problem is the availability and cost of data to develop operational measures. For example, OFCCP exists to improve the job opportunities of women, blacks, and other disadvantaged groups. Firms covered by the program are expected to make an effort to employ each group in proportion to their availability for various jobs.

Yet, the nature of availability is not easy to operationalize convincingly. Existing statistics on labor force composition may not be disaggregated by the appropriate area, occupation, or other employee characteristics to fit the conceptual specification. Developing such data, even for a sample of areas, may be prohibitively expensive.

The response to both types of problems is to employ proxy measures that provide valid indications of the variations in desired effecgs. The health effects of OSHA, for example, could be measured in the short run by exposure levels. Alternatively, the compliance rate or changes in that rate between repeated inspections may provide a proxy for treatment effects where the ultimate program effects are expected to fall outside the experimental period. In ERISA experimentation, one method would be to develop composite actuarial measures of predicted pension fund adequacy with the assistance of appropriate experts in the field, and to employ these predictors as short-run outcome measures. This procedure, however, may require substantial non-experimental analysis as a prerequisite for establishing the validity of these short-run predictors.

Experimental Period

The tradeoff between the need to provide an experimental period long enough to observe and assess valid experimental results and the need to minimize costs is an issue of concern common to all experimental designs. The time dimension involved in the regulatory process makes this tradeoff a central question in assessing the feasibility of regulatory experiments. Three aspects of this issue, which all have implications for the length of the experimental period, are discussed here.

Implementation of the treatment. A basic assumption of a classical experimental design is that the treatment stimulus is in fact established at the nominal level required by the design. If this assumption does not hold, evaluation results become ambiguous and difficult to interpret. A finding of no treatment impact may mean, for example, either that there was no response or that respondents did not perceive the treatment and thus did not respond. An essential element in the design, therefore, is to ensure that the subject of the experimentation accurately perceives the treatment in order to make a chosen behavioral response to it.

Implementing a reporting or disclosure treatment policy does not present much of a problem in this respect. The subjects can be directly informed that they are facing a new standard or requirement. However, when implementing an enforcement treatment it is relevant whether or not the firms would or should be informed of the treatment level they are exposed to. Part of the effect of the inspection effort comes from the uncertainty concerning when the inspection might occur. Disclosing information to the firm about inspection probabilities may decrease this threat effect. However, disclosing this information ensures that the treatment is perceived at the start by the subjects.

It is unclear a priori how long a training or start-up period is necessary before it can be assumed that the subject perceives and is reacting to the administered treatments. Depending on the proposed frequency of inspections and the means by which treatments are disclosed to firms, this period may last from several months to a year or more. Although the learning effects contaminate the analysis of the true effect of the treatment, analysis

of the learning effects themselves may be desirable as a source of information about the feasibility and problems inherent in full-scale implementation of the policy. Lags in the response to treatment may also appear at the end of the experiment. Extending the observation period beyond the termination of treatment policy could then provide data to be used to examine the decay in compliance effects. For example, a reasonable hypothesis that might be examined is that more intensive enforcement of OSHA safety standards has a positive impact on reducing accident rates which not only persists but gradually gets more marked over the next years even if the level of enforcement returns to lower preexperimental levels.

Temporary vs. permanent responses. The objective of a policy experiment is to provide estimates of the reposnses or effects that can be expected under a full-scale implementation. However, a fundamental limitation of the experimental strategy is that it is of limited duration and hence may not succeed in stimulating the long-run permanent responses that would be made under full implementation. This issue has been extensively explored in the contex of the negative income tax experiments. A similar problem in this context may occur when the probability of inspection/review is varied. It is not clear that firms would respond either at the same rate or in the same way to increased enforcement efforts if these efforts were perceived to be temporary. In the negative income tax experiments, the response to the 3-year experimental period was checked against subsamples of experimental subjects who were provided treatment for 5- and 20-year periods. Given that regulatory treatments increase operating costs for the participating firms, the extension of the experimental period for an experiment designed to evaluate the effects of regulation

may generate responses that create public concern and opposition, such as shifting production or closing facilities. The possibility of such long-run responses and their unintended consequences should be considered both in terms of the experimental period and the appropriate unit of analysis.

Realization of outcomes. The design of regulatory experiments is complicated by the fact that ultimate program effects may not be measured within an experimental period as short as 3 years. For example, the effect of a change in OSHA regulations or enforcement directed toward health hazards might not be known for 15 or 20 years. Similarly, the effects on pension receipts of a change in ERISA will only become apparent as the pension experience of experimental and control sample firms accumulates over 10, 15, or more years.

Program effects may lag considerably behind treatment application for other reasons. Increased enforcement efforts, for example, may result in a lengthy appeals process in which compliance and the ultimate program effects are delayed several years. Even when compliance is achieved in principle, the ultimate program effect may be postponed even further because of delays in getting new health or safety technology operational.

Time lags between treatment application and program effect increase the required experimental period and the associated cost. Longer experimental periods may also increase the problems of maintaining the cooperation of participating firms and the potential for political opposition. One set of alternatives is to focus on outcome measures that are proxies for ultimate

program effects (i.e., exposure levels as proxy for the incidence of diseases among employees). It is not clear, however, what alternatives are available if firms are faced with substantial compliance costs and respond by exercising their appeal rights.[9]

Unit of Analysis and Stratifications

The smallest unit of analysis in a regulatory experiment will normally be the firm (establishment) since it would probably not be feasible to subject a given firm to more than one experimental treatment. However, for ERISA experiments, the coverage of a pension plan is likely to be the smallest possible unit, again due to considerations of administrative feasibility. In many cases, even a larger unit of analysis may be desirable. For example, rival firms assigned to different experimental treatments may perceive a threat to their competitive positions. In these cases the unit of analysis might shift to the industry rather than the firm.

Many of the possible experiments with which we are concerned deal with variations in enforcement policies where it is important to try to estimate both direct effects (e.g., firms actually inspected under a new inspection policy) and threat effects (e.g., firms subject to the probability, but not the actuality, of such inspection) of a particular strategy. If threat effects are to be estimated, the unit of analysis must be a set of firms. Natural groupings, where firms are likely to communicate with each other thus facilitating threat effects, are industry and geographical area.

---

[9]The cost of litigating potential appeals could be considered as an associated if unintended outcome of the experimental process that would have budgetary implications for the agency. To reduce such problems, it may be necessary to focus on experiments that do not greatly increase the costs to any firms.

If industries and/or regions are the unit of analysis, the scale of the experiment will obviously have to be larger than if firms are the unit. Administratively, the larger units may sometimes be more convenient. For example, when regional offices play a large role in enforcement it may be easier to vary policies rather than within regions. Equity in the treatment of competing firms (and possible associated reductions in legal problems) suggests that, ceteris paribus, it would be best to vary experimental treatments across rather than within industries. If regions (or industries) are used as the unit of analysis, the costs of measuring the results of the experiment could be kept manageable by only analyzing data from a subset of all the firms included in the experimental design.

Our work has suggested the need for a number of potential sample stratifications in the sampling designs for the experiments, some of which are specific to one regulatory area. For example, in an experiment on the disclosure provisions of ERISA, the effects of various treatments may differ according to type of plan. Therefore, some pension plan stratifications are likely to be useful (e.g., by complexity of plan, number of people covered, whether the plan is a single or multiemployer plan, and whether the plan is collectively bargained). For OSHA, technological conditions are likely to be especially important, suggesting stratification by industry (and perhaps also by size of firm). For OFCCP, hiring rates and the availability of minorities are important, suggesting stratification by industry and labor market. For ERISA, the profitability of the firm is likely to be especially important, again suggesting stratification by industry.

Cooperation of Firms

Experimental treatments in the form of higher inspection probabilities, additional or different reporting requirements, or more specific targeting strategies put an increased burden on firms receiving the treatment. The nature of this burden and the degree of voluntary cooperation to be expected raises a number of issues with regard to both the design and the ultimate feasibility of potential experiments.[10]

Treatment burden. There are two ways in which participating firms can be made worse off because of the experiment. One type of burden occurs in experiments involving reporting or disclosure requirements, where the firm would bear the cost of administering the treatment (i.e., complying with the new standard). In these cases, it is relatively easy to minimize or eliminate these burdens by subsidizing such costs. Alternatively, in many cases involving reporting requirements, the treatment may be defined in such a way as to reduce current costs of compliance. In both cases, we do not expect the shifting of costs of administering the treatment from the firms to the agency to have any effect on the outcome measures.

A second type of burden is inherent in the nature of most enforcement strategies. In these situations the cost of compliance may be quite high. Moreover, it may not be easily borne by the agency because the treatment effect operates via the firm's response to the increased cost

---

[10] One approach to dealing with the problems created by possible lack of cooperation by firms is to use individual establishments (or divisions) within one large firm as the unit of analysis. If top management (and corresponding union officials) supported the experiment and if the firm were large enough and decentralized enough so that it had a number of autonomous subunits, then an experiment might be undertaken with little fear of legal (or political) complications.

of noncompliance.[11] If the costs of compliance are substantial, then the experiment may affect the competitive position of participating firms. One way to minimize this difficulty is to use industries (or industries within a given area) as the unit of analysis rather than individual firms (see the preceding division).[12]

Data quality. The level of cooperation forthcoming from participating firms can affect data quality in several ways. Perhaps the most serious danger is that of intentional misreporting by firms. This may be a particularly difficult problem in areas where enforcement is targeted by means of past accident or compliance rates or where financial incentives are used to further compliance. In each case, the experimental treatment creates an incentive for the firm to change its data reporting behavior. Reliance on statistics reported by the firm necessitates addressing the issue of potential reporting biases.

Accessibility to data sources is another data quality problem where the cooperation of the firm is essential. Measures of secondary treatment effects or plant characteristics may be construed as sensitive information by some firms. Such measures may also be unavailable in the detail or format required for the analysis, necessitating additional expense and/or efforts by the firms in order to obtain the data. Cooperation from firms is likely to be greater if they believe the experiment will lead to useful results and if firms are reimbursed for any administrative expenses that result from

---

[11] Subsidizing these costs would essentially change the nature of the experiment from one of studying inspection effects to one examining responses to financial incentives in the form of reduced compliance costs.

[12] To the extent that current enforcement strategies already place unequal burdens on otherwise similar firms because of area differences in enforcement staff, industry composition, or whatever, the issue of minimizing treatment cost impacts may be moot.

the experiment, including those related to data collection. (Encouraging cooperation will not be enough, however; it will also be necessary to establish careful quality control procedures including penalties for inaccurate reporting, at least in cases where there is any incentive to misreport.)

Although enlisting the cooperation of firms is important and perhaps essential to getting some types of data, the very process may generate biases in the data. One bias is that of self-selection into the sample, where the firms who basically perceive no threat from the regulations (i.e., these financially strongest, or already in compliance) are most likely to agree to cooperate with the study. A second source of bias may be generated among control firms who know they are in a study, and for that reason initiate changes in their operations or policies over the course of the experiment that they might not have undertaken otherwise (i.e., the Hawthorne effect). In this case, the estimated treatment effects would be biased toward zero. For example, an experiment to enforce OSHA health standards may require special monitoring of levels of exposure to harmful substances. As firms in the control subsample become aware of exposure levels they might independently move to reduce such levels. All experimental studies are subject to this potential source of bias, but a continuing relationship between the agency and the participating firms which extends past the experimental period makes an induced response among controls even more likely.

3. RESEARCH AND EXPERIMENTATION IN OSHA

Introduction

As stated in the OSHAct (Section 2b), the purpose of the legislation is "to assure so far as possible every working man and woman in the nation safe and healthful working conditions. . . ." The primary method used to achieve this obejctive has been the promulgation by the Secretary of Labor of mandatory safety and health standards applicable to most employers. Compliance with these standards is enforced by inspecting individual establishments and the issuance of citations and fines for noncompliance. Numerous experiments are possible that focus on the impact of these regulations and/or their enforcement on job safety and health.

First, measurement issues are discussed. This is followed by an examination of possible experimental designs and treatments, with the major emphasis on experiments that deal with enforcement of existing standards. The section concludes with an analysis of experiments involving labor-management safety committees.

Measurement of Outcomes

Before examining specific experimental possibilities, it is important to lay out possible safety and health outcome measures. In the safety area five measures seem obvious: (1) the probability of experiencing any accident; (2) the probability of experiencing an accident resulting in the loss of at least one work day; (3) average work days lost per injury; (4) the probability of experiencing a fatal accident; and (5) Worker's Compensation data collected in individual states. The first four measures are, or can

be, calculated from data collected from firms included in the annual Bureau of Labor Statistics (BLS) survey. Since as a general rule all establishments with greater than 100 employees are included in the survey, an experiment confined to this sample could utilize these outcome measures of accidents. In any experiment the first three measures could be based on an anaylsis of establishment and accident experience. However, because fatal accidents are relatively infrequent (about 5000 per year in the entire private sector), determining the effect of a treatment on this outcome would require more establishments in each treatment cell to differentiate between the impact of the treatment and random events than would be required for the other three measures. If the number of establishments in each cell required for such an analysis becomes prohibitive from a cost standpoint, it may be necessary to concentrate on the first three measures.

In addition to the BLS data, an experiment confined to a few states might be able to make use of the state Workers' Compensation (WC) information. The advantage of these data is that firms have a financial incentive to report accidents and may, therefore, be less likely to underreport accidents to avoid a particular experimental treatment. A potential problem with this data source is the difference across states in the reporting requirements. If WC accident data from several states cannot be compared, then evaluating treatments that vary across states is likely to be difficult. Nevertheless, the use of these data should not be overlooked, especially if an experiment is confined to a single state or to states with similar reporting systems.

In the health area, outcome measures remain a problem. The BLS data are generally recognized as inadequate. Although the health measures OSHA

is now trying to gather (i.e., Health Insurance Data, Workers' Compensation, Cancer File) may be useful in identifying industries or occupations that face specific hazards, the data may not yet be refined or disaggregated sufficiently to be useful in an experiment. Despite these problems, certain experiments dealing with specific enforcement strategies could attempt to deal with specific health problems where OSHA standards have already been promulgated. In such a situation the dependent variable would be exposure levels. The primary problem with this type of measure is that, at some point in the experiment, data would have to be collected for each observation.

Depending on the experimental treatment, additional intervening variables may also be measured and serve as proxies for outcomes. Possibilities include employee complaints, Workers' Compensation claims, and labor-management safety committee activities. These proxy variables are likely to be important where treatments are directed at outcomes that can only be evaluated after a considerable time period. Ideally, however, one would also like to validate these proxy measures. For example, although increased safety committee activity might be an important short-run measure of treatment effect, ultimately one would also like to determine if this activity leads to fewer accidents and/or health problems.

General Experimental Design for Enforcement Strategies

Before discussing specific enforcement strategies, it will be helpful to construct a general experimental design that could be used to evaluate various enforcement strategies. Within this general design, the treatment would be various enforcement strategies or methods.

The major advantage of such an experiment is that one can take advantage of random assignment of observations to treatments. In the OSHA area the basic enforcement treatments consist of (1) targeting inspections, primarily on certain firms, based on some criteria such as past accident experience, and (2) randomly subjecting some firms from a population of firms to varying probabilities of inspection.

Several geographic units of analysis are possible candidates for variation in enforcement activity.[13] One possibility is to use OSHA regions since current enforcement activities are coordinated through the regional offices. There are, however, two major drawbacks to this approach. First, in some regions very few federal inspections are conducted because of the predominance of state plans, making treatments in these regions very difficult or impossible to control. Second, in order to achieve sufficient variation in treatment levels most of the OSHA regions would have to be involved in the experiment. The economic and political costs of transforming an entire social program into a social experiment in this way are likely to be very high.

A less ambitious but equally valuable experiment would be to vary treatments by states instead of OSHA regions. Participation in the experiment could involve both state plan states and states directly under federal jurisdiction, or it could be confined to states falling into just one of the two categories. The cost of an experiment confined to just a few states is likely to be significantly less than a nationwide experiment.

---

[13] Using a broad unit of analysis, such as geographic areas (or industries in geographic areas), rather than having firms as the unit of analysis will permit estimation of threat as well as of direct effects (see Section 2, "Unit of Analysis and Stratification," for further discussion of this issue).

Varying treatments by state might also minimize the dependence of the experiment on OSHA accident data. If the states had comparable Workers' Compensation data, this information could be used for targeting enforcement activities. As noted earlier, this would minimize the possibility of the treatments causing underreporting of accidents by employers because of the economic incentive to file Workers' Compensation claims.

Assuming there is just one treatment, the design might look like that shown in Figure 4.

Figure 4

OSHA Experimental Design One

Geographic Area A	Geographic Area B
Treatment Group	Control Group

However, there are two problems with this approach. The first is that geographic effects may confound the evaluation of the treatment. The second relates to industry-specific effects. Although one could apply Figure 4 to either all firms in all industries in a region or confine the design simply to one industry, there are problems with both alternatives: the former is likely to be very expensive, and the latter will not allow generalization of the results to other industries.

One way of getting around each of these problems is to subject a different industry in each region to the treatment. In its simplest form this design would resemble that shown in Figure 5, where

Figure 5

OSHA Experimental Design Two

	Geographic Area A	Geographic Area B
Industry C	Treatment	Control
Industry D	Control	Treatment

Industry C in area A and industry D in area B receive the treatment. The advantage of this design is that the treatment is orthogonal with respect both to industry and to area so that the evaluation of the treatment is not contaminated by region or industry effects. Its main weakness is the assumption that there is no significant interaction effect between industry and geographic area. This problem can be reduced with a larger sample of industries and geographic areas.

This simple design, however, has only limited policy implications. Enforcement strategies or techniques can take on various levels of intensity. For example, in a targeting strategy one could subject all firms in a group for whom the accident rate is above the mean for the group to a 0.8 probability of inspection in a given period and subject other firms to a 0.1 probability. Alternatively, one could use the same two probabilities but only use the 0.8 probability for firms that have an accident rate one standard deviation above the mean for the group. The former strategy is obviously much more costly than the latter and may not yield sufficient additional benefits (lower accidents) to justify the additional cost. To be able to make these kinds of judgments from an experiment, different cells must be subject to different treatment levels. Thus, in Figure 5, industry C, area

A may receive one treatment level and industry D, area B another. However, once this is done the treatment levels are no longer orthogonal to regions and industries and the possibility of contaminated results reappears.

To eliminate industry and region effects and still allow variation in treatment levels, a more elaborate Latin Squares design can be used involving more than two areas and industries. For example, if one wanted to administer five treatment levels then one would administer the treatment to five industries in each of five geographic areas. In areas that do not receive the treatment, the industries could serve as a control group. This design is shown in Figure 6, where T is the treatment level.

Figure 6

OSHA Experimental Design Three

Industries \ Geographic Areas	1	2	3	4	5	6	7	8	9	10
1	T_1	T_2	T_3	T_4	T_5					
2	T_2	T_3	T_4	T_5	T_1		CONTROL CELLS			
3	T_3	T_4	T_5	T_1	T_2					
4	T_4	T_5	T_1	T_2	T_3					
5	T_5	T_1	T_2	T_3	T_4					

In this design the treatments are orthogonal to both region and industry, so effects can be evaluated without contamination.

Simple experimental treatments. In each cell of Figure 6 the firms would be subjected to a particular experimental treatment. The treatments could be different probabilities of being either inspected or

reinspected.[14] An experiment that varies reinspection probabilities is an attractive one. In particular, it more effectively incorporates financial incentives into the experiment because the fines for willful, repeat, or failure-to-abate violations are considerably higher than penalties for first-time serious violations during an initial inspection.

In another enforcement experiment, the treatments could be various strategies for targeting inspections on certain firms. One approach would be to develop average industry by establishment size accident data and concentrate general schedule enforcement activity primarily on those establishments whose experience is substantially worse than average. Targeting based on the performance of the individual firms not only appears fair, but will give firms a financial incentive to perform better and thus avoid the various costs associated with a higher probability of inspection. We suggest confining this treatment to firms with more than 100 employees because it avoids certain small firm problems,[15] and also ensures that the BLS accident data are available. The targeting across the different treatment groups might be based on the three different accident measures outlined earlier. An important secondary effect that would have to be evaluated is the effect of

---

[14] The Occupational Safety and Health Administration has developed a procedure for deciding which firms will be subject to a programmed inspection. Based on this strategy 95% of the programmed inspections are to occur in "high hazard" sectors of employment. Thus, the probability of an inspection varies by industry but the selection of firms within an industry and geographic area is largely random. This policy could conceivably be incorporated into one of the experimental designs outlined in Figures 1 and 2. It is different from the selection of firms for inspection based on past firm experience, which we refer to as a "targeting" strategy (see Mackenzie, 1978).

[15] In particular, accident rates in the recent past are likely to be subject to much random variation for small firms.

the experiment on reported accidents. This analysis would be very important in protecting the integrity of the accident data and determining how under-reporting affects the estimates of the treatments. The accuracy of the reporting could be validated with Workers' Compensation data and by conducting employee and union interviews. Because good firm-specific health data do not exist, safety rather than health would have to be the outcome measure of interest in this experiment.

In both the targeting and random treatment strategies, inspections in response to fatal or catastrophic accidents and employee complaints would be unaffected by the treatment schedule. For those industry-by-geographic-area cells subjected to the experiment, only the general schedule inspection program would be determined randomly.

Administratively, neither of these experiments appear unreasonable. Once the treatment levels or formulas are determined, the firms subject to inspections can be mechanically determined within the population defined for the experiment. Varying treatment levels across industries within a region or state should not be prohibitive since the content of the inspections does not change and OSHA currently varies enforcement efforts by industry within each region. The major administrative and experimental problem with both these experiments is ensuring that the distribution of inspectors across regions or states corresponds to the number of inspections required in each geographic area. The problem can probably be overcome by careful choice of the industries and geographic areas to be subjected to the experiment.

More complex variations. All the designs mentioned so far involve one treatment that is administered at several levels. Within one experiment it is possible, however (as we discussed in Section 2), to include

multiple treatments in each cell. One very attractive candidate for this type of experiment would be a treatment composed of both an initial random inspection and reinspections.

It is possible to conduct an experiment where the treatments are just an initial inspection probability or a reinspection probability. If an experiment on reinspection probabilities were conducted, however, it would also be desirable to control for the probability of an initial inspection. To control for both effects, the experiment would consist of two treatments: Xi, the probability of being randomly inspected, and Ri, the probability of being reinspected given that an initial inspection had occurred. Different combinations of Xi and Ri would form the treatment for each cell. In this experiment the cells in the design would resemble those in Figure 7.

Figure 7

OSHA Experimental Design Four

	Geographic Areas				
	1	2	3	4	5
1	X_1R_5	X_2R_1	X_3R_2	X_4R_3	X_5R_4
2	X_2R_4	X_3R_5	X_4R_1	X_5R_2	X_1R_3
3	X_3R_3	X_4R_4	X_5R_5	X_1R_1	X_2R_2
4	X_4R_2	X_5R_3	X_1R_4	X_2R_5	X_3R_1
5	X_5R_1	X_1R_2	X_2R_3	X_3R_4	X_4R_5

In region 1, industry 1, the firms' probability of being initially inspected is X_1 and, for those inspected, the probability of being reinspected is R_5.

Similar treatment combinations that include both inspection frequency and targeting strategies might also be considered in a single experiment. In

Figure 7 the Xi's could refer to different inspection frequencies and the Ri's to different targeting strategies. This experiment would allow a comparison between the effectiveness of targeting and the effectiveness of simply increasing inspection probabilities.

For accidents, the above experiment could be conducted using BLS firm-specific accident data. For health hazards, one would have to measure exposure levels as part of the experiment, which would make the handling of establishments more elaborate. In addition, because of the problems encountered in identifying health problems and their causes, the experiment should involve a population of firms with a known and serious health hazard. For example, all the firms in the population may use a common technology or chemical in the production process that is identified with a health problem in the absence of special control activities by the firm or the government.

Ideally, one would like to randomly assign firms in an industry with a known health hazard to the four treatments outlined in Figure 8.

Figure 8

OSHA Experimental Design Five

Groups	Time Periods t_1	t_2	t_3
1	Inspect and measure exposure level		Measure exposure level only
2	Inspect and measure exposure level	Reinspect	Measure exposure level only
3			Measure exposure level only
4	Measure exposure only		Measure exposure level only
	Treatment Period		Outcome Measures

In this experiment, groups 3 and 4 serve as the control groups. Comparisons of groups 1 and 2 with group 3 will give an estimate of the effect of the inspection and reinspection treatments only if the exposure levels are the same as t_1 for all groups. (Random assignment plus reasonably large sample sizes should result in little difference in exposure levels across groups.) Group 4 is included to separate out the effect of measurement from the enforcement effect. Establishments in Groups 1 and 2 may improve health conditions, not because of potential enforcement efforts but because they now know a health hazard exists. Group 4 provides a test of this measurement effect because no enforcement accompanies the measurement. At time period 3, the exposure level would have to be measured in all four groups. This measurement as well as the initial measurement in Group 4 would have to be done without either inspection or threat of investigation based on the results of the measurement. Thus, legally, this measurement treatment would have to be done by the experimenters since OSHA cannot inspect without citing violations. Ethically, it also raises problems for the experimenters since it would require that no action be taken following the measurement of hazardous levels.[16]

---

[16] In addition to the potential legal and ethical problems, it may also be impossible to convince firms that the measurement activities in Group 4 do not represent an inspection. Even if, for any of these reasons, it is decided that measurement without enforcement cannot be done, the experiment may not be completely ruled out. Groups 4 and 2 could be dropped and the experiment would involve a comparison between Groups 1 and 3 to determine the effect of initial inspections. In both of these groups, inspection as well as measurement would occur at t_3, which means that during the experimental period both groups would be subject to threat effects that might vitiate the usefulness of such an experiment. Another alternative is to inspect Groups 4 and 2 outside the experiment at t_4. There is a similar problem with this approach, however, since the expectation of inspection at t_4 is likely to affect the behavior of firms during the experimental period.

Possible Experiments Involving OSHA and Labor-Management Committees

In addition to experiments based on variations in the number, type, and targeting of OSHA inspections, possible experiments that relate to labor-management committees in the health and safety area have also been discussed. The potential advantage of a labor-management approach is that it increases the role of the parties immediately affected. Other things being equal, we suspect that procedures centering on the parties directly involved will be more efficient than those centering on the federal government, since the parties directly involved should have a more detailed understanding of the circumstances of the case.[17] On the other hand, it may be that the local parties do not have enough information at their disposal concerning accident and especially health hazards to perform effectively. Consistent with this view is the evidence that labor-management committees have frequently floundered into relative inactivity soon after they have been created.

From a policy (or experimental) point of view, the government is not likely to be able to simply require the establishment of labor-management committees. It can, however, create incentives for such committees. For example, the Interagency Task Force on Workplace Safety and Health (1978, Recommendation 8, pp. iv-15) has recently recommended that OSHA should promote innovative flexibility by increasing the availability of variances from existing standards.[18] This approach could be used as an incentive for the formation of labor-management committees. If a greater incentive

---

[17] This argument has been emphasized by Wayne L. Horvitz (1978), director of the Federal Mediation and Conciliation Service.

[18] This idea is also being investigated by Nicholas Ashford in studies he is doing for OSHA.

is needed, it might be possible to provide an exemption from all OSHA inspections (or at least from those not in response to catastrophes or complaints from affected workers) for plants which form committees that meet minimal guidelines. Such proposals could be tried as a demonstration in an individual industry, perhaps one with significant health and safety problems despite a high degree of unionization and high profit rates at most firms. Alternatively, such an approach could be tried experimentally by giving this option to half the plants in such an industry, with plants assigned to treatment or control groups randomly--perhaps with stratification by plant size and/or regions.

Several problems with experiments involving the use of labor-management committees might occur, however. First, to the extent that it takes time for such committees to get established and become effective, the results of this experimental treatment could possibly take longer to appear than for some other experimental treatments. The results might also be significantly contaminated by Hawthorne effects. Second, the incentive to form labor-management committees may not be strong enough to attract many participants. The incentive to the firm--not being subjected to OSHA inspections--could be increased by combining this experimental treatment with the treatment discussed earlier for more frequent inspections and/or greater penalties for violations. The incentive to the union (or the workers in a nonunion plant) is to have more direct influence over health and safety conditions. The strength of the incentive will depend on how confident the union feels that it is knowledgeable on these issues and how much power it would have when disagreements arise with management. This second issue is likely to be especially important.

If labor and management cannot reach an agreement, one possibility is for OSHA to mediate the dispute. This approach would probably require a major change in OSHA's responsibilities and staffing, however. Another possibility is to settle the matter through collective bargaining. Although there may not be any better alternative at the firm level, this approach risks the danger that safety and especially health problems will be neglected in the bargaining. Management can be expected to stress cost considerations, whereas the union can be expected to focus its primary attention on issues (like wage rates) that are more clear cut to its memebership.

A third problem area is whether unions are sufficiently knowledgeable on technical issues in health and safety to effectively represent their workers on a joint committee. The union's perception of its ability may affect its willingness to trade off regular OSHA protection for the joint committee approach. Even if this is not a problem, lack of knowledge may render the worker representatives ineffective. Recent suggestions for dealing with this problem include, in the short run, the preparation of short, clear, industry-specific pamphlets and the establishment of a toll-free hot line and, in the longer run, tuition-free courses for committee members and continued mail follow-up on evolving plant-specific knowledge.

Another useful approach to consider would be providing industrial hygienists to either work directly for the union or to serve as close consultants for the labor-management committees at several plants simultaneously. Providing such services makes the proposal more expensive, but it is an idea that might be worth trying in an experimental context. Thus, one version of an experimental approach to the labor-management committee issue would be to assign plants randomly to one of the following three groups: (1) those given

the incentive of no regular OSHA inspections if a labor-management committee is created to deal with health and safety issues; (2) those given this incentive plus a special subsidy (or other arrangement) to facilitate the use of knowledgeable people such as industrial hygienists; and (3) a control group where no changes are made in the present system.

Another approach to labor-management committees that has been suggested is to establish these committees at a broader level than the plant. For example, a state or regional level committee could help suggest where OSHA should inspect, use the power of bad publicity to pressure firms with poor records to do better, and hire the services of a wide variety of people--including engineers, architects, trainers, physiologists, and physicians--in an effort to provide services to firms and/or local labor-management committees. This idea has the side effect of encouraging these professionals to be more conscious of occupational health and safety issues. It could be tried experimentally, with OSHA indicating that it would be willing to subsidize the expenses of such committees in certain states or regions.

4. RESEARCH AND EXPERIMENTATION IN ERISA

Introduction

The primary goal of ERISA is to reduce the risk of workers not receiving adequate pension benefits, despite long-term participation in a firm's pension plan, by establishing funding standards, reporting requirements and regulations on information that must be provided to participants, and minimum vesting rules. Each of these standards can be evaluated for its ability to accomplish its stated goals--the receipt by participants of pension benefits.

Funding standards are enforced, in part, through the reporting requirements that specify what information must be reported to the Department of Labor. One purpose of these reports is to enable DOL to evaluate the adequacy of pension assets in meeting projected liabilities. Criteria to evaluate the adequacy of assets of nonterminated plans and to project termination probabilities have not been fully developed or tested for their validity. Research in this area would be valuable and could be accomplished through experimentations.

If funds of terminated defined benefits plans are not adequate to meet benefit liabilities, a nonprofit government corporation, the Pension Benefit Guarantee Corporation (PBGC), is required to meet the guaranteed, basic benefits to participants from its own funds. Thus, to cover expected liabilities because of plan failure, PBGC must also be able to evaluate funding adequacy and termination probabilities. The ideas presented below on DOL reporting requirements and PBGC premium experiments are similar in that the evaluation of funding adequacy is a necessary component in the design of both experiments.

Another area of research is the evaluation of the effectiveness of disclosure requirements in increasing the information on pension funds to participants. Although this is an important regulatory area, the design of an experiment is complicated by the difficulty of specifying the dependent variables--the understanding by participants of plan provisions and their subsequent actions on the basis of this understanding.

The vesting provisions of ERISA require plans to meet at least one of three minimum vesting standards. The impact of different vesting provisions could be assessed through either an experimental or nonexperimental design. The feasibility of both approaches is discussed below.

Generally, we believe that the most convenient unit of analysis for experimental research would be the pension plan, since all firm employees may not be covered by the same pensions plan. In most cases pension plans cover either all or groups of employees within a firm. In the following discussion, the plan will be assumed to be the unit of analysis. However, a critical design issue will be the treatment of those pension plans that cover several firms of a single employer. The unit may have to depend on the characteristics of each case (e.g., regional distribution of firms, separability of firm accounts).

Because ERISA is a relatively recent program, there is not an extensive body of nonexperimental research on its effects. Thus, nonexperimental research on ERISA may have a high benefit/cost ratio in and of itself, and may be required for the successful design of experimental research (e.g., the validation of outcome measures, the identification of meaningful vesting provisions). Research suggested below includes nonexperimental research ideas which should, and in some cases must, be done before any experiments are initiated.

We have discussed each of these areas with officials in the administrator's office of the Pension Welfare Benefits Program section of the Department of Labor. They indicated that research and possible experimentation in the reporting and disclosure areas would be particularly useful and policy relevant. The office is in the process of funding nonexperimental studies on pension information disclosure and the effects of different vesting requirements. The reporting area, therefore, may represent an area in which additional research would be particularly useful.

Research Possibilities

Reporting. One of the major provisions of ERISA requires all pension plans to file an annual report with DOL. This report includes an accountant's audit of the fund, an actuarial assessment of the fund's assets and liabilities, a schedule of benefits paid, leases and loans that are in default, and other information that will inform DOL of the status of the plan. The need for research in the reporting area stems from the need to further develop procedures for the enforcement of ERISA reporting provisions. An unsophisticated auditing procedure, which currently sends a large number of flagged reports for desk audit, has burdened a small audit staff. A more sophisticated computer auditing system would reduce the number of unnecessary desk audits and allow a closer audit of "true" problems. Second, a clearer definition of what variables or interrelationship among variables would define compliance is necessary so that the data reported by firms are useful when audited.

Two major advantages of research in the reporting area were mentioned at the March 1978 conference. First, the Secretary of Labor has been granted

considerable discretion by the Act to alter reporting requirements. Second, it has been tentatively reported that accounting firms involved in plan audits are generating a great deal of information beyond that actually required for reporting under ERISA. Thus, additional information beyond that required by ERISA regulations is available for some plans.

Any research on the reporting provisions of ERISA should attempt to achieve four major ends. First, an assessment of the actual reporting procedure should verify that data are in fact reported as required and that the data are accurate. Second, examination of the information reported should define key flagging variables essential to the audit procedure in order to give DOL data on which an accurate assessment can be made of the probability that a plan will be able to meet its pension commitments in the future and that funds are not being misused. Third, attempts should be made to minimize the information required and the frequency of reporting by pension plans. Finally, alternative enforcement procedures should be examined with respect to ERISA reporting regulations.

Both experimental and nonexperimental research with various reporting procedures and forms would help achieve these ends. Before any experiment is considered, a thorough analysis should be undertaken of plans that have terminated because of inadequate (or misused) funds since ERISA regulations have been in effect. Since ERISA may have been a convenient scapegoat for firms that wished to terminate plans due to other cost considerations, initial terminations under ERISA may not be representative, a consideration that could be tested empirically. Also, a distinction between terminations due to actual or threatened involvency versus other types of terminations (e.g., plan conversions or rollovers into IRAs) should be made. Since a

number of studies on plan terminations have already been done or are in progress, any new research effort could borrow heavily from these earlier findings. Examination of two sets of variables would be helpful: (1) those variables that are actually defined on ERISA reporting forms (e.g., form 5500--the annual financial report); and (2) variables that are not currently requested, but which are suspected to have high predictability values for plan failure.

Assessing the actual reporting procedure to verify that the data are in fact reported accurately could be accomplished by examining the data as reported. To avoid the bias that may occur in measuring reporting accuracy for only those plans that have failed, a random sample of plans that are in operation and presumed to be in compliance should be chosen. These plans could then be subjected to an in-depth examination by independent accountants to assess the accuracy of the information reported to DOL on past annual report forms.

Predicting the probability of plan failure or misuse of funds could be done in consultation with accountants and actuaries, using currently reported data supplemented by additional data on plan or firm characteristics collected from a sample of plans and firms. The results should yield predictive models with sets of flagging variables whose coefficients indicate the strength of various elements in predicting plan failure. Variables with weak or insignificant coefficients could then be rejected and new reporting requirements or flagging procedures drafted reflecting the findings on predictive variables and their interactions.

With this information in hand, possibilities for experimentation with new reporting procedures (including both information required and frequency of reporting) could be investigated. A possible approach would be to vary

the information required based on the models previously described to predict plan failure. Variation could also occur in the frequency of required reports. In both cases, the variation would most likely focus on ways to reduce the volume of reported information without reducing the effectiveness of the enforcement mechanisms built on the reporting provisions of the Act. A design matrix for this type of experiment is presented in Figure 9.

Figure 9

ERISA Reporting Experimental Design

Frequency of Reports		Type of Data Reported: Currently Required Data (C)	Type of Data Reported: Only Flagging Variables Required (F)
Currently Required Reporting Frequency	(C)	CC	CF
Less Frequent Reporting	(L)	LC	LF

Pension plans from various industries could be assigned to one of the four (or more, depending on the number of models to be tested) groups at random. A stratified random assignment procedure could then be used to control for other variables expected to influence pension plan failure (e.g., type of plan, complexity of plan provisions, number of people covered, whether the plan is a single or multiemployer plan, and whether the plan is collectively bargained). Alternatively, these variables could be controlled statistically.

The key problem of such a study would be determining the time frame needed for analysis. Obviously the ideal would be quite lengthy in order to allow plans to take their natural course. However, in an experimental setting an arbitrary time frame would be necessary. Another central problem would be the development of dependent variables that could serve as proxies for long-run pension plan success. To date nonexperimental research has not validated potential proxies. One possibility would be to have actuaries or accountants monitor the plans over the course of the experiment on the basis of the information reported. These same actuaries could then evaluate each plan in depth at the conclusion of the study to assess the plan's ability to meet its pension commitments in the future and to see if funds had been misused.

Experimentation with enforcement strategies over the course of the reporting experiment could also occur. This would mean adding a third dimension to the design matrix already specified. Options in this third dimension could include regular investigations as they occur now, targeting investigations according to some predesigned plan, and complaint investigations. Detailed research design on investigations and enforcement should be based on a systematic assessment of current ERISA enforcement strategies, the criteria currently used for investigation, and the penalties levied for noncompliance. At the conclusion of such an experiment, the costs of various investigation patterns on the reporting schemes could be estimated for the entire enforcement program connected with ERISA. The cost data would be valuable, but it should be kept in mind that even without it, it would be important to know if variations (most likely abbreviations) in the current reporting procedures can lead to an increased (or equal) assurance that a plan will meet its pension commitments in the future.

PBGC premium experiment. One of the controversial areas of ERISA has been the ability of Pension Benefit Guarantee Corporation premiums to cover the costs of expected plan failure. Premiums are currently a flat rate per participant across all plans. This premium is constant despite differences across plans in the adequacy of pension assets to cover liabilities in case of plan termination. The only premium differential is between multiemployer and all other plans, with the former enjoying a lower per participant premium.

The difficulty facing PBGC is that of constructing a premium schedule that will cover the PBGC-guaranteed benefits to vested participants in plans that terminate. However, failure of plans is not a calculable risk in the same sense as are other currently insured risks (death, fire, retirement, etc.), both because insufficient time has passed to assess this probability and because the causes of plan failure are not sufficiently understood (Bureau of National Affairs, 1976, p. 5).

It is known, however, that plan failures place a more severe burden on PBGC funds to the degree that plan assets will not cover guaranteed benefits. Thus, it may be a more logical policy to structure premiums such that they would serve as incentives to maintain adequate reserves. An experiment could be designed to test the effects of adopting an incentive system that links premiums to the asset sufficiency of the plan, and to indicate the cost to plans of adopting a different premium structure.

A strong rationale for such an experiment is provided by the Act itself, which allows the Corporation to adopt premiums based on the unfunded basic benefits of a plan (Sec. 4006). Such a premuim has not been adopted, in part, because these premiums have been considered insurance against a

risk with unknown probability, rather than an incentive to improve asset positions. An experiment would assess the feasibility, cost, and effect of an incentive system on the level of plan assets.

In such an experiment, plans would be randomly assigned to treatment groups in which different premium structures would be specified. Although various premium structures could be tried experimentally, for experimental purposes one sensible assumption would be to try only one new approach, with the control group facing the present PBGC premiums. Premiums would be designed to reflect the financial soundness of the plan, with those which designated as having sufficient funds paying lower per participant premiums.

Development of a premium structure will require a carefully defined and feasible method of fund adequacy. The same flagging variables suggested for the reporting experiment could be used to estimate fund adequacy. An alternative method, capable of being implemented earlier, could be based on the current PBGC regulations on estimating fund sufficiency in case of termination. Briefly, PBGC has specified five "priority categories" of benefits which funds of terminated plans must cover in successive order. Guidelines are also given for estimating plan assets which are allocated to these basic types of benefits. Asset deficiency is, then, the difference between the assets available and the present value of these five groups of benefits. An incentive structure for premiums could be structured using this system. Premiums would then be based on the present value of basic benefits that could not be covered by plan assets.

In developing such a system there is a risk that fund inadequacy, which triggers higher premiums, could lead to greater inadequacy and perhaps

hasten termination. Before higher premiums are levied, a grace period may be necessary to allow the plan to take remedial action. On the other hand, higher premiums could be allocated to a plan-specific account that would offset some of the deficiency in plan assets over the long run. In addition, provisions may have to be made to account for asset deficiencies resulting from temporary fluctuations in the securities market or changing employment conditions.

The ultimate outcome measures would be fund adequacy measured directly, the number of terminated plans in each treatment group, and the sufficiency of assets when terminated. Short-run variables of interest would be the percentage of plans taking remedial action to improve asset positions (indicated by subsequent changes in premiums), the costs of calculating fund adequacy (to either the plan or PBGC), and the fluctuations in premiums over time.

Disclosure experiment. ERISA disclosure requirements are those specifying information that must be given or made available to participants and beneficiaries of pension and welfare plans and to terminated employees. It is through disclosure that employees are made aware of plan characteristics, their rights to information about the plan and their expected benefits, and are given information which may enable them to assess the future value of their plan contributions. Participants at the March conference of project consultants felt that this was an area in which experimentation was both feasible and desirable.

The Act requires three types of reports. To beneficiaries and participants of a plan a Summary Plan Description (SPD) must be furnished. The SPD must include information on administration, provisions of a relevant

collective bargaining agreement, eligibility requirements for participating and benefit receipt, vesting requirements, circumstances leading to disqualification of participants or denial of benefits, claims and grievance procedures, and sources of financing. This description needs to be updated only every 5 years in cases of plan amendments, or every 10 years otherwise. In addition to the SPD, a summary statement from the annual report submitted to DOL describing assets and liabilities, and receipts and disbursements must be sent to each participant and beneficiary each year. Finally, terminated employees must receive a copy of the statement of their vested benefits that is filed with the Social Security Administration. Upon written request, any employee must be provided with a statement of his/her accrued and vested benefits. The only requirement of the annual report to participants, beyond the type of information to be included, is that it "be written in a manner calculated to be understood by the average plan participant, and shall be sufficiently accurate and comprehensive to reasonably apprise such participants and beneficiaries of their rights and obligations under the plan." The summary annual report to participants must only "fairly summarize the latest annual report" (ERISA, Sec. 104, b).

The outcome measures of a disclosure experiment are not obvious. The purpose of these requirements, as stated in the Labor Department Interim Regulation (1977), is to give "the participant or beneficiary an understanding of how the plan works, what benefits it provides and how to get them. It also provides basic information for making decisions on things like changing jobs or retiring." Thus, the purpose of these requirements is to immediately increase the "understanding" of expected benefits and of plan financial attributes among plan participants. This, in time, would ensure that more

participants would be eligible for pension benefits over the long run since they could make labor market decisions taking into account (and reducing) the risk of losing benefits. However, in the short run, over the likely period of an experiment, disclosures may have either a positive or negative effect on the probability of participants being vested by a given plan. For example, disclosure might increase the probability of vesting, even in the short run, by dissuading participants who are nearing their vesting eligibility date from terminating, thereby losing their eligibility. On the other hand, disclosure could increase terminations of participants who, fully informed about their own plan's provisions, realize that alternative job opportunities offer more advantageous pension plans. Likewise, the understanding of a break in service requirements may allow participants to leave jobs for short periods of time without jeopardizing vesting. Although in the long run more individuals may be vested for benefits, disclosure may result in greater short-run movements in and out of plans and, therefore, a lower percentage of participants in a single plan being vested. Given the uncertainty about the short run effects of disclosure on vesting probability, it is important to carefully define the expected short-run and long-run outcomes (or their proxies) of the disclosure requirements.

Before describing experimental possibilities with respect to disclosure, four nonexperimental studies are suggested.

1. A thorough study of who receives pension benefits and of the reasons why some do not should be undertaken. Although the current cohort of retirees is made up largely of pre-ERISA participants, a survey of these retirees would provide baseline data with which to compare future retiring

cohorts for an examination of ERISA's impact. Questions of eligibility and reasons for not being vested could be collected as part of the Current Population Survey or become a regular item in the Social Security Administration (SSA) Survey of New Beneficiaries.

2. The extent to which plans are complying with disclosure requirements should be determined. A random check of DOL files would ascertain if plans have filed SPDs. A follow-up survey of these plans' new participants and beneficiaries would check on their receipt of SPDs. Current employees could be surveyed to check on the receipt of the required annual report.

3. The extent of compliance regarding disclosure requirements to terminated employees should also be ascertained. This could be accomplished by surveying a sample of employees who have terminated employment in order to determine whether they have received the required pension information. The terminated employees could be selected from the termination reports submitted to SSA. The extent to which termination reports for all terminated employees are submitted to SSA should also be examined. It might be possible to do this by examining SSA earnings records over a period of time to determine all persons for whom earnings records stop (implying termination) and then checking to see whether termination reports on those persons have been submitted to SSA.

4. A survey of retirees that would assess the accuracy with which they were able to predict benefit eligibility, benefit amount, and, in some cases, plan failure should be undertaken. The expectation is that different methods of disclosure do have an impact on the ability of plan participants to anticipate subsequent events. In order to avoid the biases

of retrospective survey, this information may best be obtained from a longitudinal survey of plan participants near retirement age or in a sample of plans with different failure probabilities.

The treatments in disclosure experiments would relate to various requirements that would be imposed in terms of (1) the frequency of disclosure of individual status to plan participants, and (2) the types of information required to be disclosed. With respect to (2), for example, one treatment would be to require employers in the experimental group to provide detailed information on pension plan eligibility status to all participants on a systematic basis. Another treatment would be to require employers to provide only information on pension rights and instructions for procedures to use to obtain detailed information on eligibility status. Requirements governing the frequency of these disclosure could also be varied.

At least three sets of outcomes are of potential interest in assessing the effects of disclosure treatments. The primary set of outcomes, directly relating to the objectives of disclosure set forth earlier, would be the participants' understanding of their status and their subsequent mobility behavior (decision to leave, not to leave, etc.). Another outcome of interest in this regard would be the number of complaints registered by employees about the provisions or financial aspects of pension plans.

A second set of outcomes would have to do with the costs of complying with the various treatment requirements. A third set of outcomes would relate to the specific disclosure strategies that the plans adopt in response to the requirements imposed. For example: What methods do plans use to provide detailed eligibility information to participants? Do they provide information even more frequently than is required (e.g., with each paycheck)?

An example of a disclosure requirement experiment is set forth below. The experiment would test the effect of different disclosure requirements on the three types of outcomes mentioned above. At t_1, plans would be assigned

to one of two treatments. Plans in the first treatment group would be required to provide annual detailed information to each plan participant on his/her eligibility status, appeal procedures, and procedures for obtaining more information. Plans in the second treatment group would be required to provide annually to each participant only instructions on procedures for obtaining detailed information on eligibility status (i.e., the burden would be on the individual to initiate the request for detailed eligibility information). The control group would consist of firms with no requirements beyond those already imposed under ERISA.

At t_2, the three sets of outcomes would be measured. One short-run outcome measure would be the participant's understanding of his/her status measured by a standard test administered across all treatment groups. Longer-run outcomes, measured at future times, would include the probability of being vested, individual job mobility and, ultimately, benefit receipt. The cost of complying with the requirements in each treatment could be measured at t_2. Finally, the specific strategy employed to comply with the requirements is also of interest. For example, do plans in the second treatment group go beyond the minimum requirements by providing the required information more frequently than once a year? Are specific methods, such as a standard notice on every paycheck, more effective and less costly than others?

In participating in these experiments, plans could be reimbursed for costs of extra reporting requirements beyond those of the current regulations. This would reduce reluctance to participate and would elicit information on the costs (or savings) of different disclosure strategies.

Experimentation on type of disclosure should cover a variety of plans--single employer, multiemployer, or plans covering unrelated employers. In addition, the experiment should distinguish between collectively bargained

plans and others, since the employees under the former would presumably have greater awareness of plans through the bargaining process. Participation and outcomes will be influenced by the type of plan management. Thus, management type, which may determine plan efficiency, should be controlled.

Vesting. One of the issues that received considerable discussion when ERISA was being considered by Congress is that of vesting provisions of pension plans. Since many plans had long and inadequately explained vesting and service provisions, employees often found they were not eligible for a company pension upon retirement despite long service with a firm. In some cases, employees would be dismissed just prior to being vested, giving rise to the claim that many firms dismissed employees to reduce pension costs.

To deal with these problems ERISA requires plans to meet one of the following three minimal vesting requirements:

1. Ten-Year Service Rule--100% vesting at 10 years of covered service
2. Graded 15-Year Service Rule--25% vesting after 5 years with specified annual increases leading to 100% vesting after 15 years of covered service
3. Rule of 45--50% vesting after 10 years of covered service or whenever age plus such service totals 45, whichever comes first. Then the vesting must increase by 10 percentage points for each of the next 5 years to reach 100%.

Considerable interest has been expressed in the effects of alternative vesting provisions on both the receipt of pensions by workers and on job mobility by workers. For two reasons, however, it was felt that this topic was not a particularly good one for experimentation.

First, such an experiment is liable to be expensive. Since it is not likely to be possible to vary vesting provisions so that different employees within a plan are (randomly) assigned different vesting provisions, an individual pension plan would be the smallest possible unit of analysis. Moreover, either large plans or large number of small plans would be necessary in order to estimate mobility rates accurately, since mobility is a relatively uncommon event, especially among senior workers.

For an experiment to be feasible, it would probably have to hold both workers and employees harmless. The only obvious way to do so would be to experiment with shorter vesting requirements and for the government to pay the increased costs of pension benefits for which a firm becomes liable as a result of the more stringent, experimental vesting requirement. If changes in vesting do have a significant impact on mobility, then the costs to the government might be quite high and would have to be honored over a long period of time.

The cost argument is still somewhat speculative since careful estimates have not been made. However, there is another, and probably more important, argument for not doing an experiment now on alternative vesting provisions. There is currently some natural variation in vesting provisions due to the choice ERISA provides among the three vesting requirements. Therefore, it appears that the extent and effects of this natural variation should be studied before any controlled experiment is seriously considered. Although selection biases are always a potential problem in a natural experiment, some idea as to the likelihood of such biases (including both magnitude and direction) could probably be obtained by studying why plans

chose the particular requirements they did. It may be that the selection of vesting schedules is, in fact, random. Since few studies have been done on the selection by plans of vesting schedules and the effect of vesting on labor mobility, it is desirable that these studies be conducted before any experimental studies are suggested.

5. RESEARCH AND EXPERIMENTATION IN OFCCP

The Office of Federal Contract Compliance Programs (OFCCP) represents one of the federal government's most important efforts to promote equal opportunity in employment. It holds authority derived from Executive Orders 11246 and 11375, which prohibit discrimination on the basis of race, color, religion, national origin, or sex. Federal contractors and subcontractors having or seeking federal contracts of $10,000 or more are required not only to eliminate employment discrimination but also to take affirmative action to provide equal employment opportunities at all company facilities, including those not working on federal contracts.

This program has proven quite controversial. Women, minority group members, and their allies frequently argue that the enforcement procedures are too weak, whereas many firms complain that the enforcement procedures are time consuming and require too much paperwork. Some regard the affirmative action requirement as an example of reverse discrimination against white males.

Our discussion of research and experimentation in OFCCP falls in three categories. First, several general issues or problems are raised that would confront any experimental or nonexperimental researcher in the OFCCP area. Second, some specific experimental possibilities are presented. Finally, some ideas for nonexperimental research possibilities are suggested.

Experimental Issues and Problems

Three problem areas have been discussed with respect to OFCCP research. First, over how wide a unit should the experimental treatment vary? Second, what dependent variables are of primary interest? Third, how can the availability of women and minorities for particular jobs be measured?

Unit of Analysis. The smallest unit of analysis would be the individual firm, since affirmative action requriements are imposed on particular firms. In some of the experimental possibilities discussed below it will be necessary to distinguish between the direct effects of some action, such as a compliance review, and the threat effect that such an action may have. In such cases, individual firms are needed as the unit of analysis for testing direct effects, and some group of related firms (e.g., those in a particular industry or geographic area) would be necessary to test for threat effects. A group of competing firms will also be necessary for testing the effects of financial incentives. To ensure that the results of an experiment have reasonably wide applicability, it would be desirable to utilize several different industries and a variety of labor market situations. To generate maximum interest in the results (and also to avoid some of the "availability" issues), however, an experiment should focus primarily on industries that are having significant equal employment opportunity (EEO) problems.

Dependent variables. The primary dependent variables for an OFCCP experiment would be the changes in the relative employment, earnings, and wage rates of women and minorities compared with white men. More specifically, for each disadvantaged group and each job group, the following dependent variables are of interest:

> Changes in the relative employment, R=E/A, where E is the employment rate of a particular disadvantaged group (e.g., women, blacks, etc.) in a particular job group, relative to the employment of white males for that job, and A is the availability of that disadvantaged group (relative to white males) for that job group.

Specifically, the results would be measured as $\frac{\Delta R}{R}$, which represents the percentage change in relative employment of the disadvantaged group, and $\frac{\Delta R}{1-R}$, which represents the change in relative employment with respect to the gap that now exists between the disadvantaged group and the reference group, white males. Similar measures would be used for changes in relative earnings and wage rates.

In addition to these dependent variables (and weighted averages of these variables) measure of relative and absolute hiring and promotion rates would also be useful.

The problem of availability analysis. One of the most critical and difficult problems in any experiment in the OFCCP area relates to the availability rate, as in the above formulas. Much emphasis has been given to the problems inherent in availability analysis both in discussions with OFCCP officials and researchers, and in the "Preliminary Report on the Revitalization of the Federal Contract Compliance Program (OFCCP, 1977).

In most cases the data used for determining availability are contained in the report entitled "Manpower Data for Affirmative Action Programs," published by the state and local employment security agencies. Contractors may use data from alternative sources, however, to determine the availability of minorities and women for jobs requiring specific skills. The aforementioned Preliminary Report (p. 77) has noted the following problems in procuring and analyzing availability data:

- Availability data do not adequately measure the contractor's own promotable employees . . . .
- Availability data often reflect only those minorities and women currently participating in certain occupations rather than measuring those who are qualified to participate.

- Statistics are of only ancillary assistance when estimating current availability of women for nontraditional industrial jobs . . . .
- Contractor concentration on statistical evidence of availability and underutilization often divert their attention from "bottom line" results.
- The criteria for determining availability . . . do not adequately measure the availability of potential applicants or employees.

Since availability is an integral part of the initial determination of compliance, as well as the specification of dependent variables in an OFCCP experiment, its measurement is particularly important. There are many possibilities for procuring availability data. A detailed labor market survey could be conducted; however, although it should give the most accurate information on availability, it could prove very expensive. The ratio of the group in the labor force (or in the population, say, aged 16-70) would be easy to measure, but might lead to other serious problems such as court challenges if assessed penalties are based on low ratios. An average of current availability estimates of firms would also be relatively easy, but might not be very accurate. In addition, if decisions affecting firms financially were known to be based on this approach, it might lead to biased reporting. Estimates based on 1970 Census data would be fairly easy to use, but again are not likely to be considered very accurate, since the data are now out of data. It may be that other data now available to OFCCP would reduce this problem. Clearly, more thorough investigation on availability data needs and data sources is necessary. A final possibility, which avoids the issue, is to develop an experimental design that is realtively independent of how availability is measured (see Appendix B).

Experimental Possibilities

Variations in the targeting of compliance reviews. One variation in compliance review procedure and targeting that could be attempted in an experiment involves reporting requirements. For certain target industries or labor markets a new system of reporting could be developed. Initially, firms would be required to report relatively little (perhaps the standard EEO reports), with additional requirements for firms whose performance was questionable on the basis of the first review. This process could be continued, step by step, until a complete compliance review would be required for some firms. Explicit exemption from compliance review for a specified period of time could be awarded to firms with very good EEO records.

A second variation would involve varying the $1,000,000 cutoff for compulsory preaward compliance reviews, which is the only formal targeting procedure currently in effect. As part of the experimental treatment, the cutoff could be increased for some industries and the resources formerly used for compliance review could go into other enforcement activities in that industry. Comparisons could be made between the effects of compulsory reviews for such contracts and the alternative enforcement efforts. This approach could be tried either as an experimental study, with the industries for a raised cutoff selected randomly, or as a nonexperimental study, where the industries are selected judgmentally. In the latter case a before-after comparison would still be of interest.

In addition to raising the cutoff for compliance reviews, reporting requirements might be varied in other ways by size of firm and size of contract (as an alternative, or in addition, to the variations by performance

discussed above). For example, firms with over 5000 employees or over $500,000 in government contracts might be required to file annual reports on their affirmative action efforts.

Finally, since data may be more readily available on government contracts than on subcontracts, contractors might be made responsible for the affirmative action efforts of their subcontractors.

Financial incentives. Considerable interest has been expressed in the possibility of using financial incentives as a supplement to the present system based solely on penalties for those not in compliance. Under this approach firms could be financially rewarded for good performance as well as penalized for a poor showing.

Several possible financial incentive approaches have been discussed. In general, such incentives could be aimed at rewarding good past performance by conferring advantage on those seeking government contracts who have good records, and/or rewarding substantially improved performance made during the course of the contract. In rewarding past performance, the incentives might take the form of additional points for good performance in the evaluation of cost-based, negotiated procurements or discounted prices in fixed-price, negotiated procurement. In rewarding future improved performance, the incentives might take the form of additional fees in negotiated procurements and a separate bonus fee plan for formally advertised fixed-price contracts, to be paid after the contractor achieved the promised performance.

The financial incentives approach is one that appears to lend itself well to experimentation. One could visualize selecting several categories of industries or businesses--probably some with a history of equal employment difficulties in one or more discrete labor markets. Within those categories,

one would structure varying treatments which would include rewards for past performance and/or rewards for future improved performance. Within each category, the level of incentive--e.g., the number of points awarded for good past performance and the added fee or bonus profit for future improvement--could be varied. Participation in the incentives would be voluntary with the burden of proof for demonstrating good performance placed on the participating firms, subject to audit. The effects on the hiring, promotion, wage rates, and employment rates of women and minorities relative to white males could be studied. The effects on firms' costs and profits could also be examined, and one could judge the implicit costs firms assign to the regulatory compliance approach through the analysis of responses to varying levels of incentives.

Presumably, a fairly detailed formula would have to be developed to determine when and by how much a firm would gain such a competitive advantage. For both this case and the sequential reporting-targeting approach above, it appears that the availability issue must be settled before the necessary targeting and/or financial incentive schemes can be developed. In addition to assessing availability by job group, it will also be necessary to weight the firm's performance for various job groups (see Appendix B for further discussion of these issues).

The legality of experimentation with financial incentives is not clear. The DOL Solicitor's Office was not certain about (1) what "criteria" could be used (either within or in addition to the concept of the contractor's ability to perform the services requested) in developing the standards for evaluating contractor bids to perform services for the federal government,

or (2) what financial arrangements could be made with contractors--specifically, whether different arrangements could be made depending on their "record" in any area including affirmative action.

We have informally explored some of the legal, institutional, procurement, and political issues with other individuals with expertise in legal and procurement matters. From a legal standpoint, one would appear to face some difficult, but not insurmountable, issues, particularly with respect to reverse discrimination arguments. If carefully tied into the affirmative action plans, which have withstood legal challenge, the experiment should be viable. Independent of the legalities, government officials would need to review the legal structure carefully and be prepared to withstand a political reaction based on the argument that the incentives represented a shift from goals to quotas.

The financial incentive approach may induce firms to make their affirmative action goals as small as possible in order to gain the incentive payments. Also, from a procurement standpoint, there may be differences in the feasibility of incentives experimentation, depending on the type of federal contract. For example, there appears to be some precedent and fewer administrative problems with cost-based, negotiated contracts in setting up rewards for past or future performance. More difficulty may be encountered with fixed price contracts, on the other hand, and experimentation with these types of contracts may not be feasible.

Training programs. For a subset of firms whose initial compliance reviews were not favorable, ETA subsidies for on-the-job training (or some other form of training) might be made available to help firms meet their

affirmative action requirements.[19] One of the difficulties here will be to try to make the subsidy large enough so that it is likely to have some effect, but still small enough so as to not encourage firms to try to obtain the subsidy by demonstrating a poor initial EEO position. Again, detailed formulas relating initial EEO performance to amount of subsidy would appear to be useful.

Other Research Possibilities

There are several other possibilities, both experimental and non-experimental, that warrant attention in OFCCP research. One would be to focus primarily on hiring and promotion rates. This would be one way to avoid the availability issue, although it does so at the cost of ignoring factors that affect the pool of applicants or the pool from which promotions could be made.

More emphasis could also be placed on the various OFCCP enforcement weapons, such as passovers, debarments, or breach of contract suits. This approach could well be combined with the development of a standardized availability analysis.

Finally, it may be useful to utilize matched pairs in an examination of differences between firms with good and bad EEO records. In such a study a set of industry-labor market combinations might be chosen (e.g., autos-Detroit, supermarkets-Dallas, etc.), and within each industry-labor market

---

[19] Although training subsidies are likely to be most useful when a firm is hiring, subsidies for not laying off women or minorities might be tried in a recession context. This approach, of course, might present substantial legal and political problems. Layoffs of minorities and women, and their relationship to seniority provisions in collective bargaining agreements continues to be a subject of considerable controversy.

one firm with a good EEO record and one with a bad record would be selected. Comparisons between the good and bad firms could be made on a wide variety of dimensions, including their OFCCP experiences (e.g., number of compliance reviews), how EEO policies are decided and implemented within the firm (e.g., the role of special EEO staff, other personnel officials, top management, foremen, etc.), relations with the union or unions, and pressure from local community organizations or public officials. Similar comparisons could also be made across the various industries and labor markets.

6. SUMMARY AND SUGGESTED FUTURE STEPS

The objective of this study has been to provide the Department of Labor with information on the feasibility of conducting experiments to assess the effects of possible changes in three of its regulatory programs--the Occupational Safety and Health Administration (OSHA), the Employee Retirement Income Security Act (ERISA), and the Office of Federal Contract Compliance Programs (OFCCP). In this section we briefly review our major study findings and suggest several future steps that should be taken to design and implement specific pilot experiments.

Section 1 of this report discussed the distinction between a classical experiment and a natural or quasi-experimental design. A major advantage of the classical experiment is that it can substantially reduce the possibility of contamination of the results of the experiment due to "selection bias," i.e., the inability to distinguish between the effects of the experimental treatment and the effects of other factors which may covary with that treatment. An experimental design can also be helpful in that it forces the policymakers and researchers to carefully consider and specify the exact nature of the experimental treatment, resulting in more precise knowledge of what is being evaluated and its effects.

General Design Issues

Clearly one of the most important issues in the design process is the specification of the experimental treatment itself. The potential treatments must correspond with the elements that comprise the regulatory process or policy. These elements can be divided into three general categories--the substantive standard or regulation itself, the enforcement strategy used to

bring about compliance with the standard or regulation, and other activities, such as financial incentives or education programs, that can be used to supplement the enforcement of the regulation.

Enforcement activities, which generated the most interest among agency officials, researchers, and other interested parties in all three areas, can be divided into two basic components, each of which can be evaluated experimentally: (1) the effect of the level of inspection/review on compliance; and (2) the effect of different methods of allocating a given amount of enforcement resources among firms. The first aspect relates to the magnitude of the enforcement activity. The second aspect relates to the targeting of a given level of enforcement resources.

Financial incentives, of course, have long been advocated by some as a supplement or substitute for the enforcement of regulatory standards. These incentives, which have generated considerable controversy in the OSHA area, appear to be viewed more favorably in the case of OFCCP, and various experimental treatments involving federal procurement policy could be implemented in this area.

Three primary issues must be faced in specification of experimental outcomes: (1) the identification of the objectives of the program; (2) the specification of other effects or unintended consequences; and (3) the development of operational measures of these effects. With respect to the first we have found that the goals and objectives of the three regulatory programs are by no means straightforward and precise. Considerable emphasis must be given to separating the objectives into single dimension elements, and to specifying precise standards of performance. In addition, decisions will have to be made early in the experimental process about the weights that

are to be assigned to the various outcomes, presumably on the basis of their relevance to public policy, in each regulatory area. The importance of secondary effects is underscored by the lack of a single dimension response to the treatment application. The effect may be reflected not only in terms of the program objectives, but also in terms of other policy relevant responses, including effects on employer costs, inflation rates, and employment and mobility patterns.

Even if the objectives of the program and other likely effects of the treatment variables can be precisely defined, they must be correctly measured in order for an experiment to assess the effectiveness of various experimental treatments. Accurate measurement of outcome will not always be easy, however, especially when firms have an incentive to misreport (e.g., when the experimental treatment involves targeting strategies or financial incentives affected by performance), or when program objectives include such subjective concepts as the adequacy of pension funds or the availability of minority and women workers.

A fundamental issue in experimental design relates to the period of the experiment. Three aspects of the tradeoff between the need for a sufficiently long experimental period and the need to minimize costs have implications for the length of the experimental period:

1. Implementation of the treatment. A basic assumption of a classical experimental design is that the treatment stimulus is in fact established at the nominal level required by the design. An essential element in the design, therefore, is to ensure that the subject of the experimentation accurately perceives the treatment in order to ensure adequate behavioral responses to it. In addition, the analysis of the

learning effects themselves may be desirable as a source of information about the feasibility and problems inherent in full-scale implementation of the policy.

2. Temporary versus permanent responses. A fundamental limitation of the experimental strategy is that its limited duration may not succeed in stimulating the long-run permanent responses that would be made under full implementation or, conversely, that the temporary treatments, which go relatively unnoticed during the experimental period, would draw considerable public concern and opposition under full implementation. These inference issues must be considered both in terms of the experimental period and the appropriate unit of analysis.

3. Realization of outcomes. The experimental design is complicated by the fact that the ultimate program effects may not be measured within a reasonable experimental period. Not only does this require the need to focus on outcome measures that are proxies for ultimate program effects; it also increases the severity of problems caused by program effects lagging behind the application of the treatment, for example, because of lengthy appeals or other institutional delays.

We have found in our study that decisions about the appropriate unit of analysis are by no means straightforward. The smallest unit of analysis will normally be the firm (establishment), since it is not likely to be feasible to subject a given firm to more than one experimental treatment. Many of the possible experiments we suggest deal with variations in enforcement policies where it is important to try to estimate not only direct effects but also threat effects. If threat effects are to be estimated, the unit of analysis must be a set of firms. Industry and geographic area are natural groupings in this regard, since firms in the same industry or geographic area

are likely to have knowledge about the activities of other firms in the group, thus facilitating the analysis of threat effects. Equity in the treatment of competing firms also suggests that it may be desirable to vary experimental treatments across rather than within industries.

Experimental treatments in the form of higher inspection probabilities, additional reporting requirements, or more specific targeting strategies, obviously put an increased burden on firms in the treatment group. This increased burden can come in the form of additional costs of administering and measuring the treatment and its effect, and in the potential additional cost of compliance. Although some of these costs, particularly the administrative costs of the experiment, can be subsidized, many will have to be absorbed by the firm. This burden, of course, can affect the level of cooperation forthcoming from the participating firms, which can in turn affect the data quality. The most critical issue in this regard is the potential incentive for the firm to change its data reporting behavior, especially in cases where those data are used in the targeting process. Because of these reasons, eliciting feedback from representative firms will be a critical element in the early phase of an experiment, as will monitoring firms on such issues as data reporting accuracy throughout the experimental period.

Topics for Possible Experiments in OSHA, ERISA, and OFCCP

As a result of conversations with government officials, a review of the literature, and the discussion at the conference, many specific topics for possible experimental research in DOL regulatory programs have emerged. These possibilities are outlined in Sections 3-5 of this report.

With regard to OSHA, variations in targeting strategies for inspection appear to be the most appealing candidate for the experimental treatment. Other possibilities are varying the average probability of inspection and/or reinspection and providing incentives for the formation of effective labor-management committees on workplace safety and health. OSHA is a prime candidate for experimentation because of the importance of and controversy surrounding the program. Given this controversy, however, and the associated political sensitivity, it is questionable how much enthusiasm for experimentation can be generated. Especially with regard to occupational health, there will also be difficulties in obtaining good outcome measures to evaluate the effects of differences in experimental treatments.

For ERISA, the most promising candidates for experimental treatments are variations in what plan administrators are required to report to the government, what they must disclose to enrollees, and variations in PBGC premiums. The effects of variation in vesting requirements is another good research topic, though one that can probably be addressed with nonexperimental research. The main difficulty in designing any experiment in ERISA that does not require observations for many years in the future is to develop proxy measures for the adequacy of pension funds and the enrollee's probability of eventually receiving a reasonable pension income.

With regard to OFCCP, the best candidates for experimental treatments appear to be variations in the targeting of compliance reviews, possible financial incentives for government contractors who have good equal employment opportunity (EEO) records, and possible training subsidies for those with weak EEO records. A primary obstacle in attempting experiments in the OFCCP area is the need to develop good estimates of the pool of women and minorities available for various jobs. Though a most difficult problem, the availability issue is crucial for any OFCCP evaluation.

Next Steps

If the idea of possible experiments with regard to its regulatory programs continues to be of interest to the Department of Labor, the following sequence of steps should be undertaken. First, a decision must be made as to which issue or issues are most appropriate for experimental research. Second, for each experimental possibility under active consideration, each of the questions discussed in Sections 1 and 2 of this report must be addressed. Third, a small pilot experiment should be developed. Only if each of these preliminary activities is successfully completed can the final step, a full-scale experiment, be undertaken with any reasonable prospect for success.

If any experiments are to be undertaken--or even seriously considered--with regard to any of the DOL regulatory programs, the crucial first step is for ASPER, together with the regulatory agencies and perhaps also with the Secretary's office, to decide which topics have the highest priority. Until such a decision is made, further analysis of such experimental possibilities appears of little value.

If a decision is made to give serious attention to one or more experimental possibilities, then each of the design issues raised in Section 2 must be carefully examined in the context of the particular research issue to be addressed by the experiment. Careful attention must also be given to legal and political issues, especially to the cooperation to be expected from participating firms.

When it has been determined exactly which questions are of primary importance, the following design issues must be addressed:

1. The nature of the experimental treatments, including how they can be precisely operationalized

2. The dependent variables (outcome measures), including the objectives of the program, unintended consequences, and the extent to which these variables can be measured accurately
3. The unit of analysis (e.g., establishment, firms, groups of firms, states, regions, industries, etc.)
4. The length of the experimental period, including how to estimate long-term effects if only a short experimental period is chosen
5. Any necessary control (stratification) variables
6. Necessary sample sizes
7. Cost constraints

In addition to these technical issues, legal and political factors relevant to the specific experiment also require particular attention. The involvement of the Solicitor's Office and the Solicitor's representatives in the regulatory agencies will be essential. Obtaining feedback from various relevant interest groups (e.g., union and industry leaders) is also important. Obtaining voluntary cooperation from firms is necessary, not only for undertaking a full-scale experiment, but also for the detailed planning of such an experiment. The willingness of any firm to cooperate with researchers, to participate in an experiment, and to provide access to the records of the firm is likely to depend crucially on the extent and nature of the Department of Labor's involvement in the experiment. Without active involvement of DOL, it is doubtful that many firms would take the notion of controlled experimentation very seriously.

If a large-scale experiment is to be actively considered by the Department of Labor, then it is highly desirable (in addition to the work outlined above) to undertake a pilot study to determine any pitfalls that

are likely to occur in practice but may be difficult to foresee in advance. Such a pilot study can be expected not only to identify important new problems, but also to give researchers flexibility in dealing with such problems, for example, by modifying the experimental treatments or other aspects of the experimental design. Although modifications can be made relatively easily before a full-scale experiment is underway, major changes are not likely to be possible once such an experiment has begun without jeopardizing the value of the whole undertaking.

In conclusion, experimental research on OSHA, ERISA, and OFCCP has the potential of yielding valuable knowledge of how to improve the design and operation of these regulatory programs. Although this study has identified a number of possible areas for experimentation, it has also demonstrated a number of important problems that must be resolved before undertaking any such experiment. Further work on the feasibility of experimentation should be focused on a very small number of experimental possibilities, to be determined by the Department of Labor.

Appendix A

CRITIQUE OF PRESENT EVALUATION STUDIES

In this appendix we present a critique of two existing studies that we regard as particularly attentive to selectivity biases that occur in the absence of a classical experimental evaluation. Little work has been done evaluating the effects of ERISA or on the effects of OSHA on workers' health. More has been done evaluating the effects of OFCCP and the effects of OSHA on workplace accidents. In our view the best study of OFCCP's impact, at least in its attention to selectivity biases, is by Heckman and Wolpin (1976), and the best study of OSHA's impact is by Smith (forthcoming).

The Heckman-Wolpin (HW) study evaluates the effect of the OFCCP program in Chicago (using EEO data, for firms of over 100 employees) on various measures of employment status for women and minorities in 1973. The primary results are for proportion of employment represented by blacks and by women. For this measure they find that having a government contract and thus being subject to OFCCP increases the proportion of black males employed but decreases the percentage of females (especially black females) who are employed.[20] Compliance reviews are found to have little effect.

Let us consider the first finding. HW find that, for firms with no contract prior to 1973, contracts were more likely to be obtained in 1973 by firms with a higher level and growth in the proportion of black males employed. Consequently, selectivity biases appear to be at work. In particular, we'd like to know whether the higher proportion of black males leads to

---

[20] Females were not covered by OFCCP until 1972.

a greater chance of receiving a contract, ceteris paribus, or whether firms seeking contracts do more hiring of blacks. In the former case, but not in the latter, HW overestimate the effect of having a contract on black male employment.

Selection biases appear less important in the case of compliance reviews, since such reviews are not strongly related to the level or change in the employment of minorities or women. Although these findings do suggest that, at the margin, compliance reviews had little effect on Chicago in 1973, the threat of facing a compliance review and/or the need to assemble information on the employment of minorities and women in preparation for a possible review may still have had some effects.[21] Neither can the study deal with the effect of OFCCP on all firms due to its effect on the climate of opinion in the country. Although such psychological factors may be especially important with regard to questions of discrimination, it is not easy to determine how they may have been affected by any one program or event.

The Smith study deals explicitly with the impact of OSHA inspections on manufacturing industry rates. Since the costs of complying with OSHA standards are often large relative to the penalties of failing an initial inspection, we expect the threat effect of such inspections to be small. Once inspected, the penalties for failing to correct cited violations are much stiffer, however. Thus, Smith's approach of concentrating on the direct effect of inspections appears reasonable.[22]

---

[21] In contrast to selectivity effects, these threat effects cannot be estimated using an experimental design where the experimental treatment is simply receiving a compliance review. Instead, different firms have to be subject (experimentally) to different probabilities of a review (perhaps including certainty of receiving or not receiving the review).

[22] Other indirect effects may be important, however, such as those on manufacturers of new equipment, who may lose sales if they do not produce equipment that will meet OSHA standards.

When Smith initially compared changes in injury rates between 1972 and 1973 for firms that were and were not inspected in 1973, he found that OSHA inspections were associated with a larger increase (or smaller decrease) in injuries. Since "it is hard to believe that [OSHA inspections] make hazards worse," Smith interpreted these results as implying that inspections are more likely to be made where hazards are increasing--a serious selectivity bias.

To try to deal with this selectivity bias, Smith compares establishments inspected early in 1973 with those inspected later in the year. Both groups are expected to be experiencing increased hazards during the year, but only for the former group will there have been time for the inspections to have an impact. This approach appears reasonable and yields plausible results. Nevertheless, it depends on a variety of assumptions. First, it assumes no change during the year in the procedure for selecting plants to be inspected. Second, since no data are available for inspections prior to 1973, and the analysis is limited to the effects of the first inspection in the year, the procedure is biased against finding program effects if OSHA quite frequently inspects establishments with high accident rates--which may explain why Smith finds a perverse effect of OSHA inspections on the change in accident rates for large firms. Finally, the methodology makes it difficult to detect long-term effects of the inspections.

Although a classical experimental design cannot solve all problems,[23] it can deal very nicely with the selectivity problems raised in both the

---

[23] One issue that a limited scale experiment cannot address is the effect of a national program on the attitudes of firms that have little direct contact with the program.

Heckman-Wolpin and Smith studies. Whether these (and other) gains from a classical experiment are worth the costs is a decision that should be carefully evaluated on a case by case basis, however.

Appendix B

FURTHER DISCUSSION OF OFCCP ISSUES

Two issues are addressed in this appendix. First, an experimental design is presented that is relatively independent of how availability is measured. Second, the discussion of financial incentives is presented in more precise mathematical terms.

A Design That Might Avoid Measuring Availability

Assume that we have an experimental treatment (T_1), and that we can use two industries and two SMSAs for the experiment. Then we have four cells as shown below:

	SMSA A	SMSA B
Industry A	1	2
Industry B	3	4

If we apply the experimental treatment to cells on a diagonal (either 1 and 4 or 2 and 3) and have the other two cells as controls, and if we assume that there is no significant interaction effect between SMSAs and industries in determining the availability of workers by race and sex, then we should be able to determine experimental effects by looking at changes in employment ratios, earnings ratios, and relative wage rates without being too concerned with the exact availabilities by industry and SMSA.

The main weakness of this approach is the assumption that there is no significant interaction effect between SMSAs and industries in determining availability. But with a wider sample of industries and SMSA, this weakness

may be less damaging. Although some interactions will favor experimentals, others will favor controls, and with a large enough sample and a Latin square design, the net effect of interaction in biasing results toward either experimentals or controls is likely to be small.

Formula for Financial Incentives or Targeting

Industries would be picked judgmentally for this experiment, based largely on an assessment of their need to improve on equal employment opportunity and affirmative action. This assessment should include data on their performance, but should also take account of political pressures and complaints. To keep the analysis as simple as possible, we shall also assume that the industries have local rather than national product markets. The experimental treatment would be given in some areas (SMSAs), whereas others would serve as a control group.

Targeting could be proportional and financial incentives inversely proportional to the following:

$$\sum_{i,j} (X_{ij} - \bar{X}_j)^2 W_{ij}$$

where X_{ij} = the percentage of earnings in the ith job group going to the jth demographic group (e.g., women, blacks)

$\bar{X}_j$ = the percentage of group j among all workers in the civilian labor force of the SMSA

W_{ij} = a weight reflecting the importance assigned to the firm's EEO performance in the ith job for the jth group.

The quadratic formula assumes that we're especially concerned with cases where the employment percentages are far from the SMSA average: